Machine learning bias correction and downscaling of urban heatwave temperature predictions from kilometre to hectometre scale

Received: 24 November 2023 | Revised: 21 March 2024 | Accepted: 17 April 2024

DOI: 10.1002/met.2200

Meteorological Applications
Science and Technology for Weather and Climate
Open Access
RMetS

RESEARCH ARTICLE

Machine learning bias correction and downscaling of urban heatwave temperature predictions from kilometre to hectometre scale

Lewis P. Blunn[1] | **Flynn Ames**[2] | **Hannah L. Croad**[2] | **Adam Gainford**[2] | **Ieuan Higgs**[2] | **Mathew Lipson**[3] | **Chun Hay Brian Lo**[2]

[1]MetOffice@Reading, University of Reading, Reading, UK

[2]Department of Meteorology, University of Reading, Reading, UK

[3]Bureau of Meteorology, Canberra, Australia

Correspondence
Lewis P. Blunn, MetOffice@Reading, University of Reading, Reading RG6 7BE, UK.
Email: lewis.blunn@metoffice.gov.uk

Funding information
Met Office Weather and Climate Science for Service Partnership (WCSSP); National Centre for Earth Observation, Grant/Award Number: PR140015; Science and Technology Facilities, Grant/Award Number: ST/W507763/1; SCENARIO NERC Doctoral Training Partnership, Grant/Award Number: NE/S007261/1

Abstract

The urban heat island (UHI) effect exacerbates near-surface air temperature (T) extremes in cities, with negative impacts for human health, building energy consumption and infrastructure. Using conventional weather models, it is both difficult and computationally expensive to simulate the complex processes controlling neighbourhood-scale variation of T. We use machine learning (ML) to bias correct and downscale T predictions made by the Met Office operational regional forecast model (UKV) to 100 m horizontal grid length over London, UK. A set of ML models (random forest, XGBoost, multiplayer perceptron) are trained using citizen weather station observations and UKV variables from eight heatwaves, along with high-resolution land cover data. The ML models improve the T mean absolute error (MAE) by up to 0.12°C (11%) relative to the UKV. They also improve the UHI diurnal and spatial representation, reducing the UHI profile MAE from 0.64°C (UKV) to 0.15°C. A multiple linear regression performs almost as well as the ML models in terms of T MAE, but cannot match the UHI bias correction performance of the ML models, only reducing the UHI profile MAE to 0.49°C. UKV latent heat flux is found to be the most important predictor of T bias. It is demonstrated that including more heatwaves and observation sites in training would reduce overfitting and improve ML model performance.

KEYWORDS

crowdsourced data, land cover, machine learning, numerical weather prediction, urban heat island

1 | INTRODUCTION

Weather and climate models can be used to inform decision makers about future overheating hazards and in the design of adaptive responses to climate change (Nazarian et al., 2022). The prediction of near-surface air temperature (T) within urban areas is of particular importance as cities are centres of population, commerce and infrastructure. However, modelling T within cities is uniquely challenging because weather conditions are affected by

anthropogenic emissions of heat, moisture and aerosols and small-scale interactions with heterogeneous land cover and urban structures (Oke et al., 2017). Due to computational limitations current numerical weather prediction (NWP) and global climate models (GCMs) typically operate at horizontal grid lengths (Δ) of order 1 and 10 km, respectively, and therefore cannot resolve micro-scale ($\sim$100 m) T variations (Barlow, 2014; Oke et al., 2017).

Using machine learning (ML) post-processing, rather than moving to smaller grid length numerical models, is an appealing way of obtaining sub-neighbourhood scale ($\lesssim$1 km) scale T predictions, since ML is faster and less computationally expensive. For example, going from $\Delta = 1$ km to $\Delta = 100$ m with the UK Met Office (UKMO) Unified Model (MetUM) results in $\sim 10^4$ increase in computational expense, whereas increasing the output grid resolution for ML models incurs little added computational expense. Despite advances in NWP model resolution (Ronda et al., 2017), land surface models (Grimmond et al., 2010, 2011; Lipson et al., 2023) and the surface description data provided to them (Masson et al., 2020), urban NWP T errors are still typically 1–2°C (Bohnenstengel et al., 2011; Ronda et al., 2017; Schoetter et al., 2020). It is possible that ML post-processing can be used to learn systematic NWP biases and correct them while downscaling to higher resolutions with little additional cost.

ML already has many applications in weather and climate and is a rapidly advancing field (Chase et al., 2022; Rolnick et al., 2022). It has been used to emulate expensive model parametrizations (Gettelman et al., 2021; Meyer, Grimmond, et al., 2022; Meyer, Hogan, et al., 2022; Rasp et al., 2018), nowcast precipitation (Espeholt et al., 2022; Kaae Sønderby et al., 2020; Ravuri et al., 2021; Shi et al., 2017), make daily to seasonal forecasts (Ham et al., 2019; Keisler, 2022; Lam et al., 2022; Lopez-Gomez et al., 2023; Pathak et al., 2022; Rasp & Thuerey, 2021; Weyn et al., 2021), downscale forecasts (Harris et al., 2022; Stengel et al., 2020), inform climate change mitigation (Milojevic-Dupont & Creutzig, 2021) and utilize citizen weather station (CWS) observations in bias correction of urban climate predictions (Brousse et al., 2023).

In order to improve our understanding of the relationship between urban surface characteristics and the thermal urban climate, several studies have exploited ML to generate high-resolution T maps (Alonso & Renard, 2020; Chen et al., 2022; Chen et al., 2023; dos Santos, 2020; Lyu et al., 2022; Straub et al., 2019; Venter et al., 2020; Vulova et al., 2020; Wang et al., 2023; Yu et al., 2020; Zumwald et al., 2021). These studies generally use a combination of T and remotely sensed land surface temperature observations and information describing the urban surface to train ML models. Fewer studies include information from NWP or GCMs in ML models to improve urban T forecasts. Exceptions include Cho et al. (2020) who bias corrected $\Delta = 1.5$ km NWP predictions of next day maximum (T_{max}) and minimum (T_{min}) T in Seoul, South Korea, using previous day T observations, NWP, and positional and topographic variables. Also, Wu et al. (2021) used T and dew point T (T_d) from $\Delta = 2.5$ km regional climate and $\Delta = 250$ m NWP models along with land cover and morphology data to train several convolutional neural network—generative adversarial network fusion ML models. The ML models input with $\Delta = 2.5$ km regional climate model T and T_d were able to emulate the high-resolution NWP T and T_d on a $\Delta = 250$ m grid (i.e., downscale).

ML approaches can be thought of as being 'hard', 'medium' and 'soft' when they replace, improve and emulate traditional meteorological models, respectively (Chantry et al., 2021). Enforcing physical principles and conservation laws in ML is difficult and is not currently common practice (Kashinath et al., 2021). 'Medium' ML post-processing approaches (as employed in our study) do not result in the ML feeding back on the driving NWP. Any unphysical relationships that the ML models learn do not influence the NWP, so the ML T predictions are perturbations about a physically consistent state.

A common issue when investigating the spatio-temporal variability of T in urban areas, and in the development and evaluation of models, is the sparsity and representativity of urban observations (Hahn et al., 2022; Mitchell & Fry, 2024; Muller et al., 2013). Due to cost, maintenance difficulties, problems in obtaining installation permission and the desire to meet World Meteorological Organization observation standards (e.g., that sites 'should be well away from trees, buildings, walls or other obstructions'; WMO, 2018), long-term urban observations tend to be made at airports and in parks. However, it is necessary to obtain T observations at locations covering a wide range of urban surface characteristics, since T is strongly dependent on them (Oke et al., 2017).

Crowdsourced observations, in particular CWSs, are an attractive source of urban T observations for national meteorological services (Garcia-Marti et al., 2023; Hahn et al., 2022; Mitchell & Fry, 2024; van Beekvelt et al., 2024). CWSs are often high-density covering many urban surface characteristics, each site typically has months to years of observations, and they are low cost to national meteorological services as they do not purchase or maintain the instruments. Crowdsourced observations have been used in thermal comfort assessment (Nazarian et al., 2021), urban climate studies (Droste et al., 2017; Feichtinger et al., 2020; Fenner et al., 2017; Potgieter et al., 2021), mesoscale model evaluation (Hammerberg

et al., 2018) and producing sub-kilometre urban T maps (Venter et al., 2020; Vulova et al., 2020; Zumwald et al., 2021). CWS T observations must undergo thorough quality control (QC) procedures, since thermometers often have inadequate radiation shielding and ventilation, missing metadata and inappropriate positioning such as next to walls (Bell et al., 2015; Fenner et al., 2021; Meier et al., 2017). The UKMO has developed the Weather Observations Website (WOW; Kirk et al., 2021), which is a cloud-based platform where CWS observations can be uploaded. WOW observations are already utilized in producing gridded 'mesoanalyses' of variables such as pressure over the UK for operational nowcasting of extreme precipitation (Clark et al., 2018).

To the authors' best knowledge, this is the first study utilizing ML to both bias correct and downscale urban T predictions from NWP, using a combination of T observations and urban surface description data. Bias correction and downscaling of $\Delta = 1.5$ km Met Office operational regional model (UKV) T forecasts to $\Delta = 100$ m is achieved using the 10 m resolution World Cover land cover dataset (Zanaga et al., 2021) and open-access CWS T observations (WOW). We aim to address the question of whether this 'medium' ML approach can provide improved T predictions over conventional operational NWP. The study focuses on eight heatwave events that occurred between 2019 and 2021 in London, United Kingdom. The article is structured as follows. Section 2 describes the methodology, including the ML workflow. Section 3 presents the ML model results and discusses the factors (e.g., ML model configuration and CWS data) influencing ML model performance. The study is concluded in Section 4. Appendix A contains additional figures and a table with the details of the predictors and hyperparameters used in each ML model configuration.

2 | METHODS

2.1 | Case studies

The study region is 0.75°W–0.45°E (≈110 km) and 51.1°N–51.9°N (≈90 km; Figure 1), which contains the Greater London area (≈ 50 by 50 km) and surrounding rural area. The study focuses on eight heatwave events (28–30 June 2019 [0], 21–28 July 2019 [572], 23–29 August 2019 [320], 23–27 June 2020 [545], 30 July 2020–1 August 2020 [213], 5–15 August 2020 [1486], 16–23 July 2021 [408], 6–9 September 2021 [365]) defined by Public Health England (PHE) (PHE, 2019, 2020, 2021; 49 days total). Numbers in brackets are the PHE-estimated heat stress-related excess mortalities in the 65+ age group

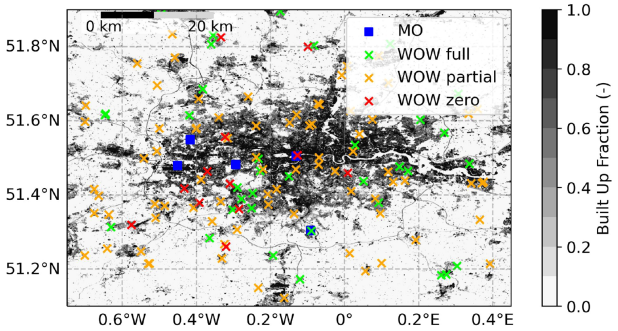

FIGURE 1 World Cover (Zanaga et al., 2021) built-up fraction aggregated to 100 m grid length in the study region of Greater London. The WOW (citizen weather station) locations with no ('zero') data for all heatwaves, partial data coverage for at least one heatwave and full data coverage across all heatwaves are represented with red, orange and green crosses, respectively. The locations of professionally maintained (MO) sites are represented by blue squares. The projection is Plate Carrée.

during each heatwave. The PHE definition of a heatwave day is when a UKMO > Level 1 heatwave alert occurs (where > Level 1 is specified by regionally varying T threshold exceedances, see supplementary material of Green et al. (2016)), or where the mean Central England T (Met Office Hadley Centre, 2024) is >20°C on the day, previous day and following day (Green et al., 2016). To incorporate meteorological conditions surrounding the heatwaves and increase the amount of data, 1 day before and after each heatwave is included in each case study (65 days total).

2.2 | ML workflow

In this section, the ML workflow (Figure 2) is described. First the data were prepared: the WOW data were quality controlled (see Section 2.3.2), the World Cover land cover (see Section 2.3.3) was aggregated to $\Delta = 100$ m, the land cover fractions 1, 5 and 25 km upstream of each 100 m grid cell were calculated (see Section 2.3.3), the UKV (see Section 2.3.1) variables were linearly interpolated to the same 100 m grid and land cover and UKV variable time-series were extracted at the grid point nearest the WOW and professionally maintained (MO) observation sites (see Section 2.3.2).

The target (i.e., the quantity the ML models are trying to predict) is the hourly difference between the UKV T and WOW T (i.e., the bias), rather than the hourly WOW T, since it gives better predictions (see Section SM-2.1 discussion of configuration group 7 for details). Hence, the ML T prediction is calculated by subtracting the ML bias prediction from the UKV T.

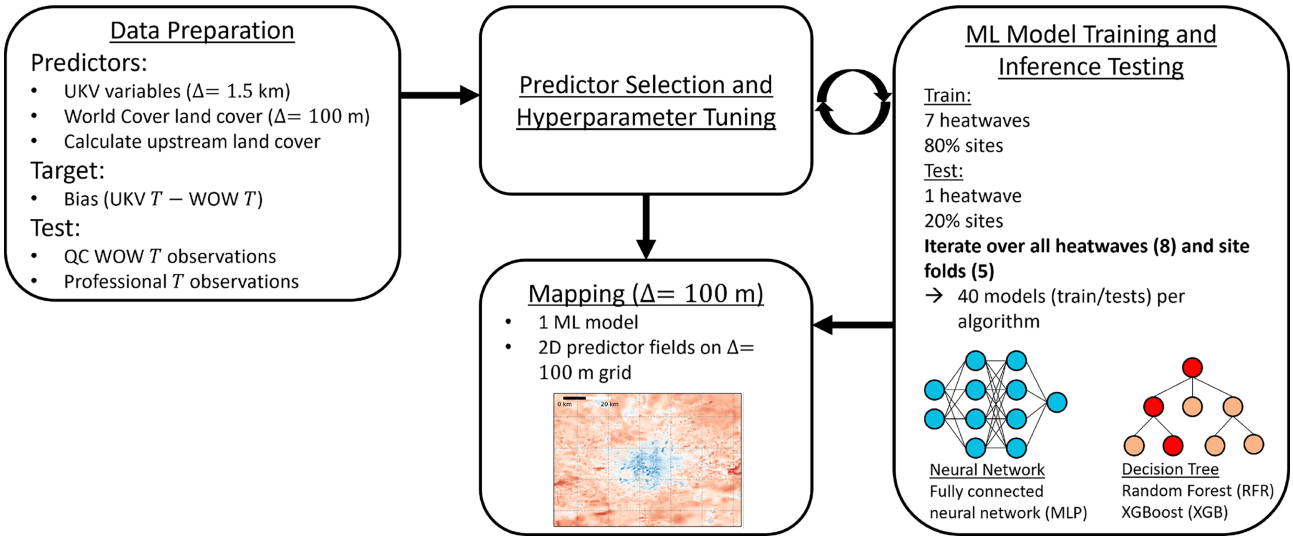

FIGURE 2 Schematic illustrating the ML workflow used in this study.

The available UKV predictors are 1.5 m air temperature (T), net shortwave radiation (K^*), cloud cover fraction below 2000 m (CL), 10 m wind speed (WS), 1.5 m relative humidity (RH), latent heat flux (Q_E), sensible heat flux (Q_H), ground heat flux (Q_{soil}), soil moisture in the top 10 cm soil layer (SM) and hour of day (HoD). The available land cover predictors are grassland ($GL\alpha$), tree cover ($TC\alpha$), built up ($BU\alpha$) and permanent water ($PW\alpha$) fraction, where α can represent the land cover at each 100 m grid point ($\alpha \equiv p1$), and 1 ($\alpha \equiv 1$), 5 ($\alpha \equiv 5$) and 25 ($\alpha \equiv 25$) km upstream of each grid point. The difference between the World Cover and UKV land cover ($\alpha \equiv p1diff$) at each grid point and 1 km upstream ($\alpha \equiv 1diff$) are also available. Some studies use building morphology information as predictors (Wang et al., 2023), but since the WOW sites are largely at open low-rise, open midrise or vegetated locations (Demuzere et al., 2022, 2024), and because the leading influence on urban land surface–atmosphere interactions is land cover (Grimmond et al., 2010), we limit the land surface description predictors in this proof of concept study to land cover.

The predictors and target were standardized using the standard deviations and mean averages of the training data. The sensitivity of ML model performance to different ML model configurations (i.e., different combinations of predictors and hyperparameter choices) is investigated (see Section 3.1). This involved an iterative process where following ML model training and inference testing, predictors and hyperparameters were updated, and so on. The ML algorithms are described in Section 2.4.

It is the intention that the ML models should be able to bias correct and downscale at all locations within the study region, for any future heatwave within current climate conditions. Therefore, the observation sites and heatwaves used to test the ML model performance should not be included in training. In cross-validation, one heatwave is used for testing and the other seven are used in training. This is done eight times, using a different heatwave for testing each time. Using this approach, predicting T for the test heatwave is analogous to producing an operational T forecast, since the test period (analogous to the future) is unseen in training. It is desired that the error averaged over all test sites should be representative of the error averaged over all grid points. This can be achieved with a random train/test split of the WOW sites (assuming that WOW sites are randomly located). A 5-fold cross-validation across WOW sites is performed for each test heatwave, where WOW sites are split randomly into 5-folds, 4 of which are used in training (80%), and the other which is used to test (20%). This is done five times, using a different fold for testing each time. Hence, for each ML algorithm, 40 ($=8\times5$) ML models are generated and inference tested. For each model, the mean absolute error is calculated as

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\, T_{p,i} - T_{o,i} \,|\,, \qquad (1)$$

where n is the number of samples (from for all sites and times) used in testing the model, T_p is the predicted T and T_o is the observed T. A single MAE from the resulting 40 MAEs is calculated as a weighted average of the case study length and number of sites in each inference test. Finally, to make maps, a representative ML model is chosen and run at each 100 m spaced grid point, using 2D predictor fields. Note than when calculating the MAE

for the UKV, to make a fair comparison, samples are restricted to those available in testing the ML models.

2.3 | Data

2.3.1 | Numerical weather prediction

The NWP to be downscaled and bias corrected is the UK variable resolution (UKV) deterministic limited area forecast model (Tang et al., 2013), which has fixed 1.5 km horizontal grid length over the study region. The UKV is run operationally by the UKMO to produce forecasts out to 120 h with a new forecast being initiated every hour (i.e., hourly 'cycling'). A global model with 10 km horizontal grid length and 6 hourly cycling provides lateral boundary conditions to the UKV. The UKV initial conditions are obtained by combining the previous UKV forecast with observations through a 4D-Var data assimilation system (Milan et al., 2020; Rawlins et al., 2007; which has time- and space-dependent treatment of forecast errors), to obtain a best estimate of the Earth's atmosphere and surface states. We use the first 6 h from each UKV forecast starting at 03, 09, 15 and 21 UTC to create a continuous hourly time series for each case study. The UKV is a configuration of the MetUM that solves fully compressible, non-hydrostatic, deep-atmosphere dynamics with a semi-implicit semi-Lagrangian numerical scheme (Davies et al., 2005; Wood et al., 2014). The land surface model is the Joint UK Land Environment Simulator (JULES), which has eight non-urban tiles each representing surface exchange for a different land cover class (Best et al., 2011; Clark et al., 2011). The urban module within JULES is MORUSES. It represents urban areas with a 2D infinite street canyon geometry that accounts for the height and separation of buildings, with the overall effect being imparted through separate canyon and a roof tiles (Bohnenstengel et al., 2011; Porson et al., 2010). Anthropogenic heat emissions vary between $16.7-26.2$ W m^{-2} for different months of the year based on 1995–2003 UK energy consumption (DUKES, 2003), are fixed diurnally and are down weighted based on the built-up fraction in each grid cell.

2.3.2 | CWS and professional observations

Within the study region, there are 133 WOW sites with T data on at least 1 day during the study period (locations shown in Figure 1). The WOW observations were downloaded from the UKMO internal system and can be accessed externally in one-site, 1-month chunks (Met Office, 2024). Observations have varying temporal frequency but are processed to hourly frequency by using the timestamp on the hour or linearly interpolating to the hour using the nearest time samples. Observations with lower than hourly frequency are set to null.

CWS observations require careful QC as discussed in Section 1. The simple QC steps and threshold values developed here are designed to remove obvious erroneous data. Since there are 8 case studies (heatwaves) and never more than 105 sites per case study, the data were examined manually to ensure robust and satisfactory filtering. As an example, timeseries for all sites during 21–28 July 2019 before and after QC are shown in Supplementary Material (SM) Figure SM-1.1. In the following steps, each site is considered separately:

1. Remove from case study if $< 50\%$ data coverage.
2. Remove from case study if $> 5\%$ of values are outside $T_{med} \pm 3 \times IQR$ (where T_{med} and IQR are the median and interquartile range T of all sites, respectively).
3. Remove values that are outside $T_{med} \pm 5 \times IQR$.

Step 1 removes sites where the CWS was not installed, was interfered with or had technical issues during the case study. Step 2 removes sites that are consistently unlike other sites, either due to having unrealistic values or poor CWS placement (e.g., indoors). Step 3 removes extreme values which are typically individual spikes. Remaining data for each WOW site were visually inspected and appeared physically reasonable.

Prior to QC, there was 49.8% data coverage over all sites and times. After QC, there is 46.0% data coverage (i.e., 3.8% of data is made null). One hundred fifteen sites have data available for at least one case study, and 35 sites have data available for all eight case studies. 46.0% data coverage for 135 sites over 65 days with hourly frequency equates to 97.3×10^3 T data points. Sites with zero, partial and full coverage are shown in Figure 1.

Figure 3 shows the mean average (T_{av}), maximum (T_{max}) and minimum (T_{min}) near-surface air temperature for the UKV versus WOW observations at each site. The UKV values are taken from the grid cell nearest to each WOW site. WOW sites with partial and full coverage are included. UKV and WOW T_{av}, T_{max} and T_{min} were determined by calculating their daily values followed by averaging over all heatwave days for which WOW observations were available. T_{av} and T_{max} are generally higher for WOW than the UKV (i.e., most points fall below the red 1:1 line), particularly at higher temperatures. T_{max} on average is 1.1°C higher for WOW than the UKV and can be up to 4°C higher. However, T_{min} is on average lower for WOW than the UKV by 0.3°C.

There are five UKMO maintained ('professional') sites available within the study region with hourly temporal resolution (Met Office, 2022). These are plotted as

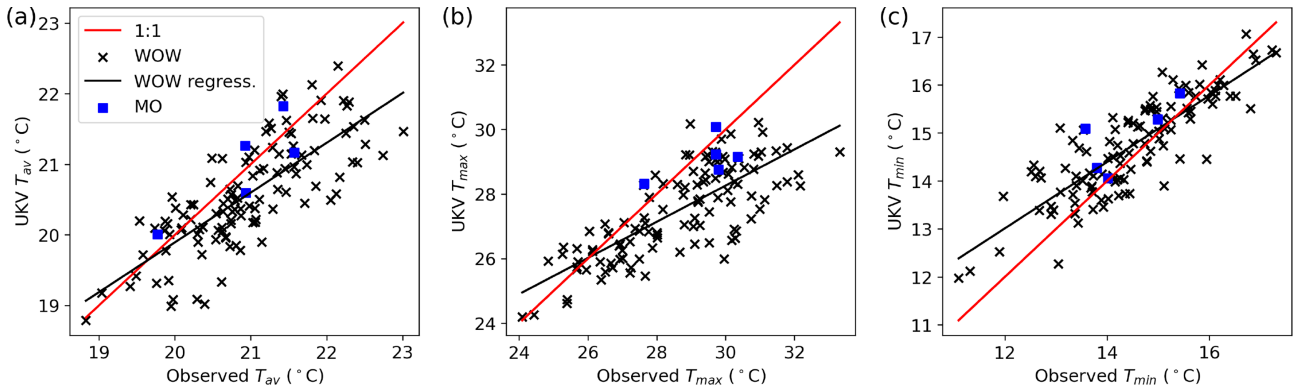

FIGURE 3 A comparison of WOW and MO observations with UKV model data at the same grid point. Daily (a) average near-surface air temperature (T_{av}), (b) maximum near-surface air temperature (T_{max}) and (c) minimum near-surface air temperature (T_{min}) calculated as an average over case study days. The black line is a least squares linear regression to the WOW points.

blue squares in Figure 3. Their T_{av} and T_{max} are generally less scattered and closer to the UKV (i.e., fall close to the 1:1 line). The reasons are likely 3-fold: (i) of the professionally maintained (MO) observations 3 are at airports and 2 are in parks, which have less variable local-scale settings compared to typical WOW sites settings, (ii) unlike WOW observations, the MO observations meet World Meteorological Organization standards for siting well away from obstructions, so are less affected by micro-climate (e.g., complex building geometry) influences and (iii) four out of the five MO sites are assimilated into the UKV initial conditions improving UKV prediction at those sites. Given the WOW T_{av} and T_{max} are increasingly higher than the UKV and MO at high temperatures, it is possible they exhibit a warm bias due to insufficient radiation shielding and/or ventilation, consistent with other CWS studies (Bell et al., 2015; Cornes et al., 2020; Fenner et al., 2017; Fenner et al., 2021; Meier et al., 2017). See Section 3.6 for discussion on implications for ML model performance. It can be seen in Figure SM-1.1 that some sites (e.g., 14, 23, 28, 49, 51 and 71) have higher than the median T near the middle of the day, and that the sites that consistently have the most extreme values are removed by the QC (i.e., sites 49 and 71). While an additional more stringent QC step could be applied to T near the middle of the day, the approach is not taken, since there is a trade-off between removing data influenced by radiation and losing real information on local-scale T hot spots.

Assessment of WOW data quality on a site by site basis is challenging because, aside from latitude and longitude, no other metadata is consistently available. T adjustments for example associated with the height of the thermometer above the ground cannot be made. Therefore, in this work, WOW observation sites are assumed to have been taken 1.5 m above the ground to enable comparison with the UKV 1.5 m air temperature predictions. An additional issue, when developing CWS QC techniques for urban areas, is that T variations of several degrees can develop between local climate zones (LCZs) that are of order 1 km in scale (Stewart & Oke, 2012). Unless several observation sites exist per km^2, it is difficult to assess whether an observation is an outlier based on other nearby observation sites. Also, unlike some CWS datasets (Netatmo, 2023), where T observations are made using a standardized thermometer type, the thermometer at each WOW site can be of variable type and quality, with potential T biases of several degrees under high global radiation levels (Bell et al., 2015). This makes QC even more challenging. Cornes et al. (2020) developed a QC technique aimed at removing shortwave radiation-related T biases from WOW observations in the Netherlands. They used WOW T, professionally measured ('background') rural T and downwelling shortwave radiation, in a generalized additive mixed model to obtain bias-corrected WOW T. A limitation of this technique is that urban and local-scale effects can be incorrectly modelled as radiation effects, hence removing the urban and local-scale signals. However, the bias-corrected T results demonstrated a qualitatively plausible diurnal representation of the urban heat island (UHI). Development of such a complex bias correction model for the UK is out of scope for this proof of concept ML study. The WOW observations are used as truth although it is acknowledged that data quality issues will be present.

2.3.3 | Land cover

The 10 m resolution World Cover (Zanaga et al., 2021) class-based land cover dataset is used for downscaling the

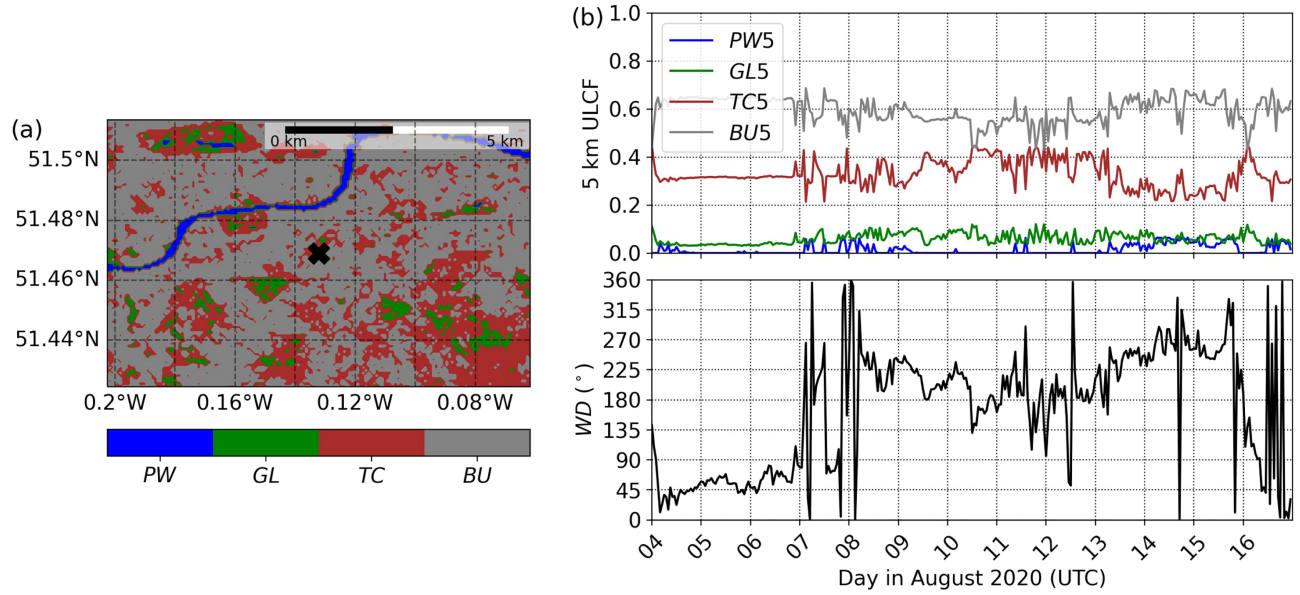

FIGURE 4 (a) World Cover (Zanaga et al., 2021) land cover data (aggregated to 100 m grid length) surrounding an example WOW observation site (black cross) in Clapham (Central London), illustrating the dominant land cover classes: built up (*BU*, grey), tree cover (*TC*, red), grassland (*GL*, green) and permanent water (*PW*, blue). The projection is Plate Carrée. (b) The 5 km upstream land cover fraction (ULCF) (top) and UKV wind direction (bottom) at hourly time frequency for the same site on 4–16 August 2020.

UKV. The dataset is aggregated to 100 m grid length so that each 100 m grid cell is comprised of fractions of each land cover class. The land cover classes considered are built up, tree cover, grassland and permanent water, since these all have greater than 0.01 land cover fraction when land cover is averaged across WOW sites.

Land cover influences T locally but can also have non-local influence via advection (Brousse et al., 2022). In addition to accounting for local-scale effects using the 100 m land cover fractions, we account for non-local impacts by calculating land cover fractions 1, 5 and 25 km upstream for each point in the 100 m grid as follows: (i) the wind direction is linearly interpolated to the 100 m grid using the nearest UKV grid points, (ii) at each 100 m grid point, the mean average land cover fraction within eight 45° upstream wind sectors ($337.5^\circ - 22.5^\circ$, $22.5^\circ - 67.5^\circ$, etc.) is calculated and (iii) a weighted linear combination of two wind sectors (based on the wind direction) is used to calculate the final upstream land cover fraction. Note that (iii) assumes the local UKV wind is representative of the upstream wind direction. 1, 5 and 25 km are chosen to be broadly representative of a lower bound on the neighbourhood scale, an upper bound on the neighbourhood scale and the city scale, respectively. In future work, more complex methods of calculating the T source area of each site could be considered, for example, accounting for the effects of atmospheric stability, wind direction changing with upstream distance and turbulent exchange between the roughness sublayer and the overlying atmosphere. Figure 4a shows the dominant land cover classes in the area surrounding a WOW site in Clapham (Central London), and Figure 4b shows the 5 km upstream land cover fractions calculated at the site for 4–16 August 2020. It can be seen that when the wind is westerly, there are larger permanent water and built-up fractions (e.g., 13–15 August 2020).

The UKV uses the 25 m resolution 1990 Institute of Terrestrial Ecology (ITE) land cover classification dataset (Bunce et al., 1990). The dataset is over 30 years old so significant land cover changes have occurred since its generation, and the dataset does not benefit from modern remote sensing techniques. For use in the UKV, the ITE land cover is aggregated to the 1.5 km grid and the land cover classes become fractional. Here, the gridded ITE land cover is linearly interpolated to the 100 m grid, and the land cover fraction difference with World Cover is calculated for built up, tree cover, grassland and permanent water. This is done to provide information on where UKV bias correction is most likely required.

2.4 | ML algorithms

Three ML algorithms are used for supervised regression in this study: random forest (RFR; Breiman, 2001; Ho, 1995), XGBoost (XGB; Friedman, 2001) and multilayer perceptron (MLP; Gardner & Dorling, 1998). They

are chosen to encompass decision tree (RFR, XGB) and neural network (MLP) type ML algorithms. All three are capable of learning non-linear relationships, which is beneficial for predicting T in urban areas, due to the non-linear nature of this system.

RFR is an ensemble learning method where each decision tree is an ensemble member, and in regression mode, the final prediction is the average of the predictions from all of the decision trees. The module RandomForestRegressor from Python package sklearn version 1.1.3 is used (Scikit-learn Developers, 2023b, 2023c). XGB is also an ensemble learning method, but unlike RFR where decision trees are independent of one another, the decision trees are trained sequentially, and in such a way that the model outcomes are weighted towards the direction in which the loss function decreases the fastest. The module XGBRegressor from Python package xgboost version 1.7.1 is used (XGBoost Developers, 2023a, 2023b). MLP is the most common type of feedforward artificial neural network. It has an input layer, at least one hidden layer, and an output layer. Each node in a layer is connected to every node in the following layer so that it is 'fully connected'. At the end of each iteration, a cost function is minimized by updating the weights associated with the node connections. We use the module Dense from Python package tensorflow version 2.9.1 (TensorFlow Developers, 2023a, 2023b) with rectified linear unit (ReLU) activation function for all but the last hidden layer that has linear activation function. In compilation, the Adam stochastic gradient descent optimization method is used.

We also use multiple linear regression (MLR) as a benchmark for the ML algorithms. MLR is a statistical technique for modelling linear relationships between the predictors and the target, and therefore cannot model multivariate and non-linear relationships like RFR, XGB and MLP. We use the module LinearRegression from Python package sklearn version 1.1.3 (Scikit-learn Developers, 2023a, 2023c).

3 | RESULTS AND DISCUSSION

3.1 | ML model performance sensitivity to predictors and hyperparameters

This section presents the main conclusions from predictor and hyperparameter sensitivity investigations (for an expanded version with more detailed discussion and statistics, please see Section SM-2). The ML model configurations can broadly be split into seven groups. The naming convention of each configuration is $X-Y-Z$ where X is the configuration group number, Y describes the hyperparameters and Z is a distinguishing feature (see Table A1 for more details). The MAE with and without 5-fold cross-validation is presented for all configurations in Figures SM-2.1a and b, respectively. The discussion herein refers to results from 5-fold cross-validation unless otherwise stated.

- In configuration group 1, the ML models were first tested with UKV T as the only predictor, and then each remaining UKV predictor was tested in turn. ML models trained using all of the UKV predictors gave the greatest improvements over the UKV, indicating the importance of multivariate relationships. Henceforth, all UKV predictors are included in each ML model configuration.
- In configuration group 2, land cover predictors (binned into 0.2 fraction intervals to prevent overfitting) are also used to train ML models. It was found that certain combinations of land cover predictors show some ability to improve performance, but using all 100 m land cover predictors at once generally degrades the results. One possible reason that the land cover predictors give less improvement than might be expected is that the ITE land cover dataset and the JULES surface scheme in the UKV already represent the influence of land cover spatial variability on T well. Another possible reason is that the relationship between 100 m scale land cover variation and WOW T is dominated by thermometer quality and thermometer placement differences between sites. To reduce the parameter space, the following ML model configurations are trained using all UKV predictors and built-up fraction predictors only (100 m, upstream, and the difference between ITE and World Cover).
- In configuration group 3, it is found that binning the built-up fraction improves RFR and XGB model performance versus not binning. The reduced MAE of the ML models trained with the binned built-up fraction indicates that overfitting, where the ML models are learning relationships from the training data that do not generalize well to previously unseen data (i.e., when tested 'out of sample'), has been reduced. Also, not using built-up fraction predictors (vs. using binned built-up fraction predictors) improves ML model performance for RFR and XGB, whereas using binned built-up fraction improves MLP model performance (consistent with MLP being the ML algorithm with best built-up fraction downscaling results in Section 3.4).
- In configuration group 4, it is found that larger hyperparameters degrade model performance, again indicative of overfitting. Neural network overfitting prevention (regularization) techniques were also

investigated in combination with larger hyperparameters, which significantly reduced MAE. However, these ML model configurations still did not give better results than using smaller hyperparameters and binning as in configuration 3.
- In configuration group 5, a wide range of 'different' hyperparameters are investigated, but these generally degrade ML model performance compared to the best-performing models in configuration 2.
- In configuration group 6, it is demonstrated that using no land cover improves model performance (compared to including land cover) when using large hyperparameters (even when binned), indicating that land cover overfitting occurs with large hyperparameters.
- In configuration group 7, it is demonstrated that using the absolute WOW T as the target, rather than the difference between UKV T and WOW T, degrades model performance. This is likely because there is a very large correlation between UKV T and WOW T, which causes training to focus too heavily on their relationship, such that it does not identify other target-predictor relationships.

The single best-performing configurations for RFR, XGB and MLP are $2-\mathrm{SHP}-PWp1$, $2-\mathrm{SHP}-BU5$ and $5-\mathrm{DHP}-e$, respectively, where $SHP=$ small hyperparameters, $PWp1=$ permanent water fraction at each 100 m grid point, $BU5=$ 5 km upstream built-up fraction and $\mathrm{DHP}-e=$ a set of different hyperparameters (see Table A1). All give MAE and root mean square error (RMSE) reductions of 0.12 (11%) and 0.18 (11 − 12%)°C compared to the UKV, respectively. The UKV has MAE and RMSE of 1.12°C and 1.55°C, respectively. Brousse et al. (2023) conducted WRF model (Skamarock et al., 2018) simulations during the summer of 2018 and evaluated their T prediction performance using Netatmo observations (Netatmo, 2023) across southeast England. Across all sites (urban and rural), MAE and RMSE were 1.8°C and 2.3°C, respectively (their Table 3). Their ML bias correction technique reduced MAE and RMSE by 0.32 (17%) and 0.29 (13%)°C, respectively. They therefore achieved slightly better percentage reduction in errors than in our study, but this could be due to their WRF simulations having larger biases than the operational UKV output used in this study. Furthermore, objective comparison of ML-based post-processing techniques is challenging since studies often use different regions, time periods, NWP models and CWS datasets (Wang et al., 2023).

Without 5-fold cross-validation (i.e., when tested 'in sample'), the best-performing configurations for RFR, XGB and MLP are $4-\mathrm{LHP}-\mathrm{C}$, $4-\mathrm{LHP}-\mathrm{C}$ and $4-\mathrm{LHP}-\mathrm{Drop}$, respectively (Figure SM-2.1b). $LHP=$ large hyperparameters, $C=$ control and $Drop=$ drop out layers regularization (Keras Developers, 2023). The MAE improvements of 0.27, 0.26 and 0.19°C, respectively, are approximately double that from 5-fold cross-validation (i.e., when tested 'out of sample'). This is because when the ML models are tested at sites used in training, their performance benefits from having learnt characteristics specific to the sites. Such ML models should be used when the aim is to make predictions at observation sites, rather than when making T maps. In the latter case, ML models must generalize to unseen locations, and so overfitting degrades performance.

Although the ML model configurations trained with built-up fraction are not the best-performing in terms of MAE, including them is crucial in demonstrating their potential in downscaling T to hectometre scale. Hence, in the coming sections, we also examine the $3-\mathrm{SHP}-\mathrm{C}^b$, $4-\mathrm{LHP}-\mathrm{C}$ and $4-\mathrm{LHP}-\mathrm{C}^b$ configurations. Examination of feature importance was performed for these configurations using RFR. It was found that Q_E has by far the highest importance in each case (Figures SM-2.3 and SM-2.4), consistent with it having a strong control on T in urban areas. This suggests that improving the representation of Q_E in the UKV could have large benefits for UKV T predictions. Q_E being the most important predictor is physically consistent, since spatially Q_E is anti-correlated with Q_H and T, and can vary a lot in urban areas due to large vegetation fraction heterogeneity (Oke et al., 2017). On average HoD, RH and WS are the next most important predictors, although the exact importance varies by configuration (Figure SM-2.1a).

A MLR model was trained using the $3-\mathrm{SHP}-\mathrm{C}^b$ predictors, to investigate whether a simple statistical model (with linear relationships) could perform as well as the more complex ML models. The ML models (all configuration $3-\mathrm{SHP}-\mathrm{C}^b$) achieved a 0.11°C reduction in MAE on average compared to the UKV, whereas the MLR achieved a 0.09°C reduction. Hence, using only linear relationships, one can achieve approximately 80% of the MAE reduction obtained by the ML models.

3.2 | Heatwave variability of ML model performance

The performance variability over the different heatwaves for the ML models (RFR, XGB and MLP) is investigated. Configuration $3-\mathrm{SHP}-\mathrm{C}^b$ is chosen since it contains the built-up fraction predictors and is one of the better performing configurations for the ML models (see Figure SM-2.1a). Figure 5 shows the difference in MAE between the UKV and each ML model averaged over all case studies (central segment) and for each heatwave

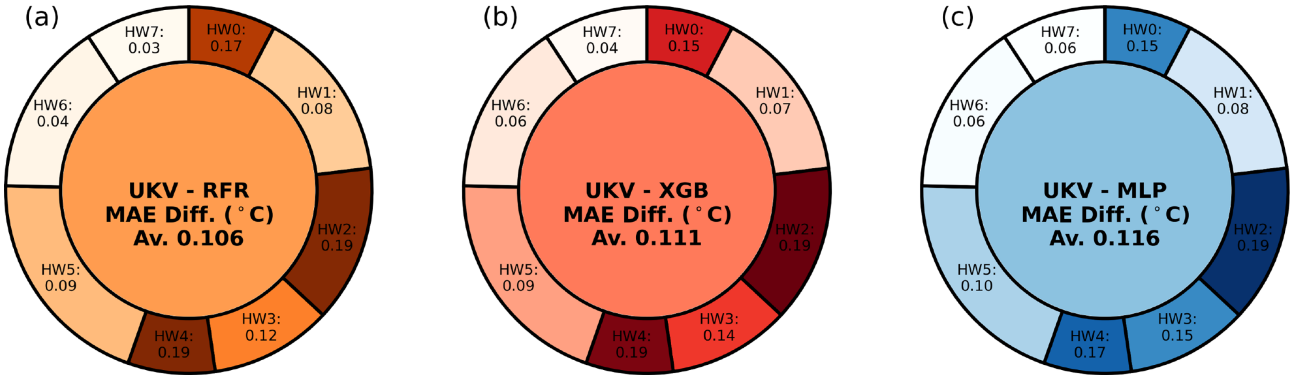

FIGURE 5 Difference in mean absolute error (MAE, relative to WOW observations—taken here as the 'truth value') between the UKV and each ML model ((a) random forest (RFR), (b) XGBoost (XGB) and (c) multilayer perceptron (MLP)) averaged over all case study days (central segment) and each heatwave (outer segments). Positive MAE indicates where an ML model outperforms the UKV in predicting T at WOW sites. Darker shading denotes a larger magnitude MAE difference between an ML model and the UKV. The arc length of each segment is representative of the heatwave length. The results for all ML models are from configuration $3-\mathrm{SHP}-\mathrm{C}^b$.

(outer segments). The MAE improvements for heatwaves range between $0.03-0.19$, $0.04-0.19$ and $0.06-0.19$°C for RFR, XGB and MLP, respectively. This demonstrates the importance of testing on multiple time periods (i.e., heatwaves) that are separated long enough in time so that observation sites are under the influence of different weather systems, otherwise, different conclusions on the ML model performance can be reached.

3.3 | Diurnal temperature and UHI bias correction

Here the ability of the ML models to bias correct the UKV diurnal T and UHI predictions is investigated. Figure 6 shows the composite all-site, all-period mean diurnal T (left y-axis) and MAE (right y-axis) for ML models (configuration $3-\mathrm{SHP}-\mathrm{C}^b$), MLR and the UKV. The MAE of the predicted diurnal T profiles (i.e., calculated after compositing) is denoted as MAEp (not plotted). When evaluated at the WOW sites (Figure 6a), the RFR, XGB, MLP, MLR and UKV profiles have MAEp of 0.11, 0.11, 0.22, 0.23 and 0.67°C, respectively. This equates to a $3-6$ fold MAEp improvement for the ML algorithms over the UKV. The ML models capture the average behaviour of the WOW sites during the heatwaves $\approx 0.52^{\circ}$C better than the UKV. Configuration $1-\mathrm{SHP}-\mathit{HoD}$ that only includes UKV T and hour of day predictors gives diurnal profiles (not shown) that are almost identical to those from $3-\mathrm{SHP}-\mathrm{C}^b$, demonstrating that the site average T temporal variation can be learnt by these predictors alone. The simple MLR model performed well for the composite diurnal profile (with comparable MAEp to MLP), which is consistent with simpler models (e.g., ML models with configuration $1-\mathrm{SHP}-\mathit{HoD}$) being able to bias correct the composite diurnal profile.

For all models, the MAEs are lowest during the evening and night and largest during the morning and afternoon (Figure 6a). During the evening and night, the ML models and UKV have similar MAEs, but during the morning and afternoon, the ML models have lower MAEs. Also, the MAEs generally increase over 6-h periods ending at 3, 9, 15 and 21 UTC, due to larger errors for longer UKV forecast lead times. However, the MAE generally increases less for the ML models than the UKV during each 6-h period, demonstrating their bias correction capability. The performance of the ML models and UKV at professionally maintained (MO) observation sites (Figure 6b) is discussed in Section 3.6.

The choice of ML model configuration has a strong influence on UHI bias correction. Figure 7 shows the average difference between T at the urban and vegetated sites for the ML models, UKV and WOW with figure panels showing different ML configurations. Sites are classed as urban when built-up fraction >0.5 and vegetated when the sum of grassland and tree cover fractions are >0.5. WOW observations show little difference in mean T between vegetated and urbanized areas at midday, but at night, more urbanized areas are up to 2.5°C warmer. The ML models offset the tendency of the UKV to have too high T at the urban sites relative to the vegetated sites in the afternoon and evening. Best ML results are obtained with large hyperparameter values and all built-up fraction predictors (see $4-\mathrm{LHP}-\mathrm{C}$ Figure 7a and $4-\mathrm{LHP}-\mathrm{C}^b$ Figure 7c). The MAEp of the $4-\mathrm{LHP}-\mathrm{C}$ RFR, XGB and MLP diurnal profiles with respect to the WOW profile are 0.19, 0.15 and 0.20°C, respectively, which is a large improvement over the UKV that already has a low MAEp of 0.64°C.

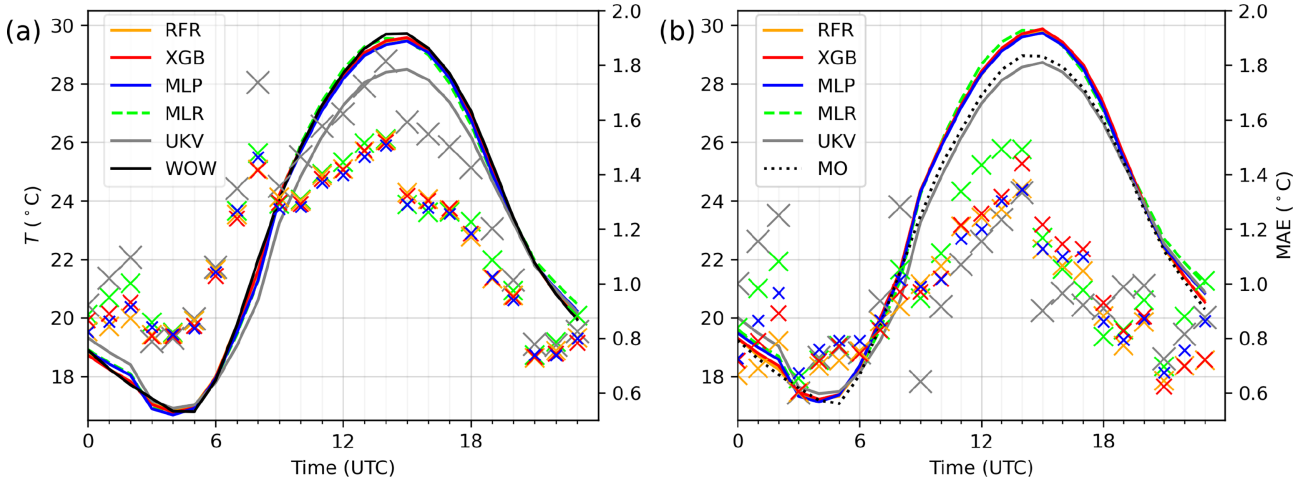

FIGURE 6 Diurnal near-surface air temperature (T; left y-axes) and mean absolute error (MAE; taken relative to WOW observations—right y-axes) composites calculated over heatwave days and sites for the ML models ($3-\text{SHP}-\text{C}^b$ random forest (RFR), XGBoost (XGB) and multilayer perceptron (MLP)), multiple linear regression (MLR), the UKV and observations at (a) WOW (citizen weather station) sites and (b) MO (professionally maintained) sites. Solid lines relate to T and cross markers correspond to MAE, with legend colours indicating data source. The cross markers have different sizes for readability when they overlap.

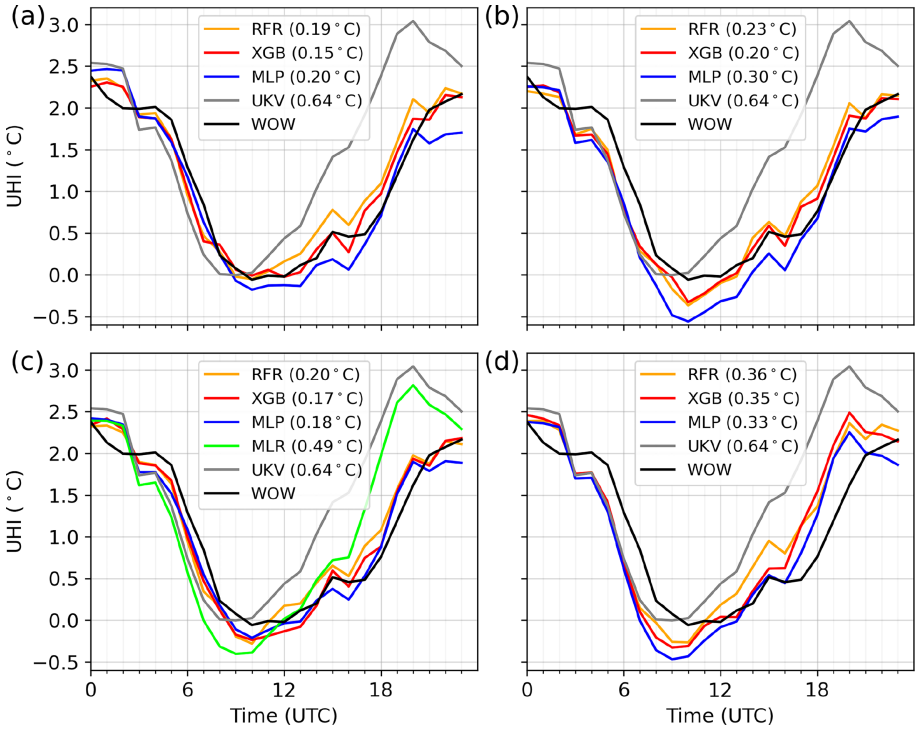

FIGURE 7 Mean urban heat island (UHI), defined here as the difference between the average T at the urban and vegetated sites for the ML models (random forest (RFR), XGBoost (XGB) and multilayer perceptron (MLP)), UKV and WOW. Sites are classed as urban when built-up fraction >0.5 and vegetated when the sum of grassland and tree cover fractions are >0.5, based on the 100 m grid cell each site is in. (a) $4-\text{LHP}-\text{C}$, (b) $6-\text{LHP}-\text{NoLC}$, (c) $4-\text{LHP}-\text{C}^b$ and (d) $3-\text{SHP}-\text{C}^b$. Multiple linear regression (MLR) is also shown in (c). Values in the legends correspond to MAEp—the mean absolute error of the model profiles compared to the WOW profile.

For $6-\text{LHP}-\text{NoLC}$, which has the same configuration as $4-\text{LHP}-\text{C}$ except without built-up fraction predictors, the MAEp is on average 0.06°C poorer across ML models compared to $4-\text{LHP}-\text{C}$ (Figure 7b). Therefore, the built-up fraction predictors provide further improvements to the ML UHI predictions. When the same configuration as $4-\text{LHP}-\text{C}$ is used, but with binned built-up fraction predictors (i.e., $4-\text{LHP}-\text{C}^b$), the performance is comparable to $4-\text{LHP}-\text{C}$ (Figure 7c), so binning the built-up fraction predictors does not degrade UHI prediction. When small hyperparameters and binning of the built-up fraction are used ($3-\text{SHP}-\text{C}^b$), the poorest results (compared to the previous three configurations) are obtained (Figure 7d). Therefore, large hyperparameters help the ML models learn the UHI behaviour.

ML models with larger hyperparameters ($4-\text{LHP}-\text{C}$ compared to $3-\text{SHP}-\text{C}^b$) have improved UHI, but degraded T MAE (see Figure SM-2.1a), which suggests there is a trade-off. Large hyperparameter values enable the UHI to be learnt, but result in overfitting degrading

the T MAE (see Section 3.1). When large hyperparameters are used overfitting is not only due to land cover predictors. This can be seen from configuration 6 – LHP – NoLC, which has large hyperparameter values and no land cover predictors. When compared to 2 – SHP – C, which has the same predictors but smaller hyperparameter values, the MAE is poorer (see Figure SM-2.1a). The exact causes of the overfitting will be the subject of future investigations.

MLR performed almost as well as the ML models in terms of MAE (see Section 3.1) and the diurnal profiles (as discussed above). However, it can be seen from Figure 7c that MLR has poorer UHI bias correction capability than the ML models, particularly in the evening. MLR uses the same predictors as 4 – LHP – C^b and 3 – SHP – C^b ML model configurations. Compared to the best-performing ML model from those two configurations (4 – LHP – C^b XGB with MAEp = 0.17°C), MLR has approximately three times as large UHI MAEp. Hence, the simple MLR model does not perform as well as the ML models in capturing the UHI, consistent with the UHI requiring complex relationships to be learnt.

3.4 | Downscaling

In Section 3.3, configuration 4 – LHP – C gave the best UHI bias correction, demonstrating the benefit from the built-up fraction predictors, so it will be used here to investigate ML model downscaling of the UKV. Figure 8 shows a T difference map (100 m grid length) between 4 – LHP – C MLP and the UKV over London at 18:00 UTC, 25 August 2019. MLP tends to make the UKV cooler in the city and warmer in the rural surroundings in the early evening. This means that urban and rural T have been brought closer together, consistent with bias correction of the UKV towards the WOW observations at 18:00 UTC (Figure 7). Hence, the ML improves the spatial representation of the UHI.

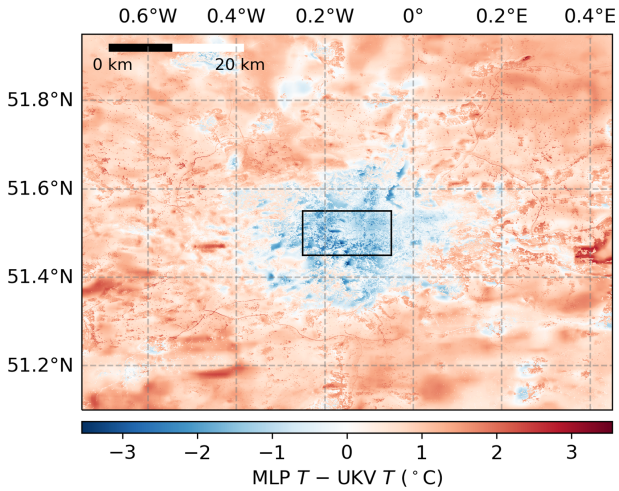

FIGURE 8 Map showing the difference between 4 – LHP – C MLP and UKV near-surface air temperature predictions at 100 m grid length over London at 18:00 UTC, 25 August 2019. Red denotes where the MLP ML model predicts higher T than the UKV. The black rectangle is the region shown in Figure 9. The projection is Plate Carrée.

Figure 9 shows (a) 4 – LHP – C MLP T, (b) UKV T and (c) 100 m built-up fraction over a smaller region (with location represented by a black rectangle in Figure 8). The downscaling results in regions of low built-up fraction having generally lower T, consistent qualitatively with what is expected during the evening in urban areas with significant vegetation cover. Also, it can be seen that large parks (characterized by low 100 m built-up fraction) tend to have low T in their south-east portions. This correlates with low 1 km upstream built-up fraction (Figure 9d) and is physically consistent with cool air in the north-east of the parks being advected by north-easterly winds (Figure 9e) south-east across the parks. The T spatial patterns in parks do not correlate exactly with 1 km upstream built-up fraction or any other individual predictor. It is therefore likely that multivariate relationships that give physically plausible behaviour are being learnt. Exploring such behaviours (e.g., the relationships between land cover predictors, and their relationships with latent heat flux (Figure 9f)) will be the topic of future investigation.

MLP is chosen because RFR and XGB demonstrate more erratic behaviour (not shown), with seemingly random grid point to grid point fluctuations. MLP behaviour can also be difficult to interpret, for example, grid cells along the river Thames typically have very low urban fraction and are expected to be cooler than the surroundings. However, the river Thames is generally cooler and warmer compared to the surroundings in the west and east of Figure 9a, respectively. Also, unexpected sharp 100 m scale warm–cold–warm patterns sometimes occur at the edge of sharp land cover boundaries, for example, parks. These patterns do not appear in any of the predictors. Although most predictors vary smoothly at 100 m scale, it might be that multivariate relationships are being learnt that have sharp tipping points, leading to sharp spatial gradients. Understanding and improving the relationship between ML model T and predictors will be the subject of future work.

3.5 | Influence of training data on ML model performance

To determine whether ML model performance can be improved if more WOW sites were available, the number of sites included in ML training is increased from 18 to 92 (i.e., 115 sites with 5-fold cross-validation; Figure 10a).

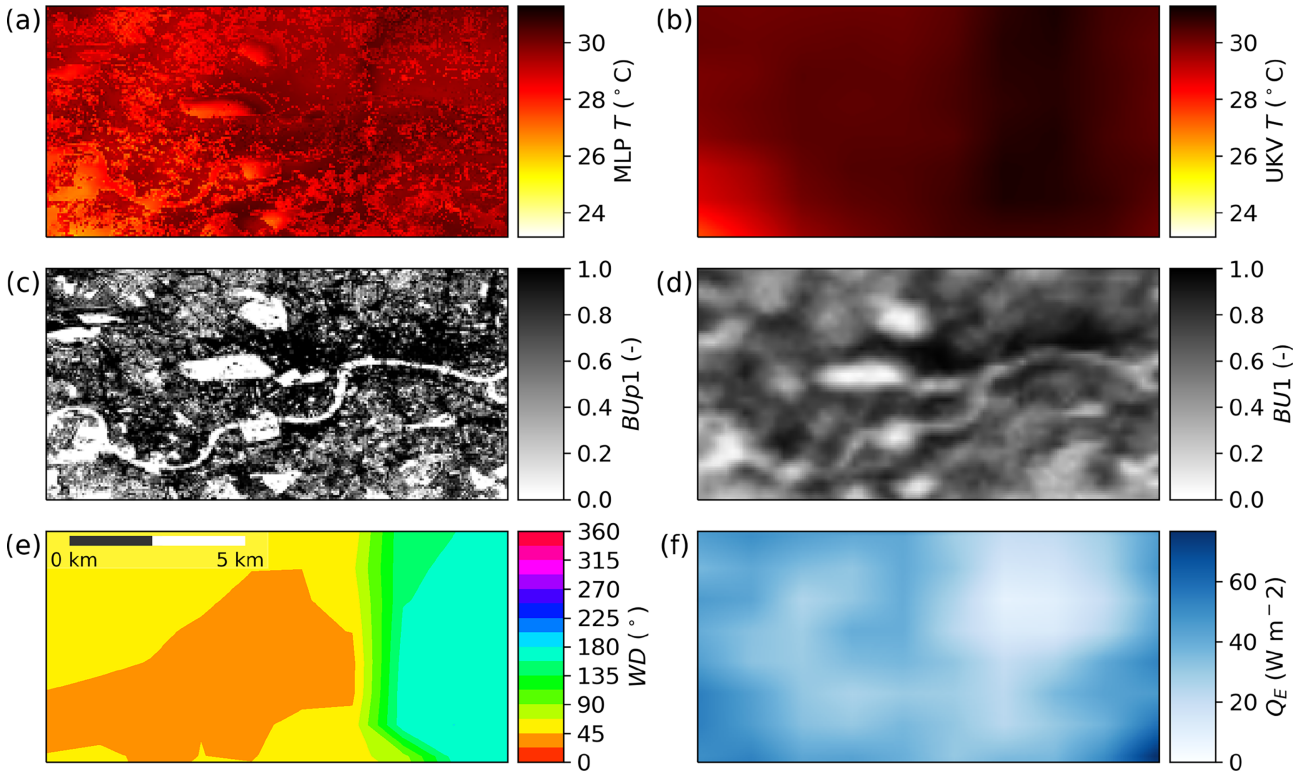

FIGURE 9 A 100 m grid length maps at 18:00 UTC, 25 August 2019 of (a) ML model (4 – LHP – C MLP) near-surface air temperature (T), (b, e and f) UKV T, wind direction (WD), and latent heat flux (Q_E) (linearly interpolated to 100 m), (c) 100 m built-up fraction ($BUp1$) and (d) 1 km upstream built-up fraction ($BU1$). The area shown in these maps corresponds to the black rectangle in Figure 8. The projection is Plate Carrée.

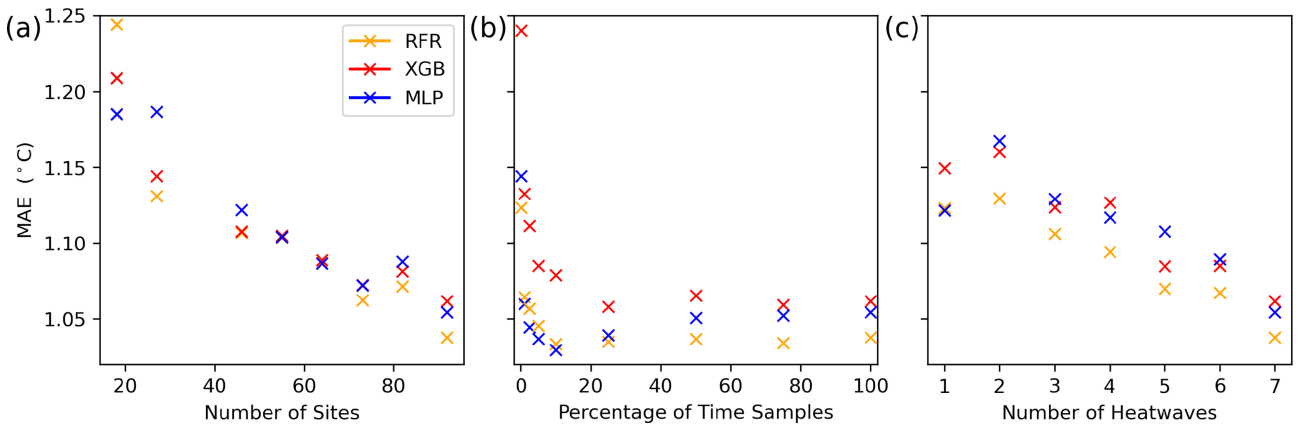

FIGURE 10 Influence on mean absolute error (MAE) (relative to WOW observations) of (a) the number of sites, (b) percentage of data points included (at random across time) and (c) number of heatwaves included in training of the 4 – LHP – C ML models.

Configuration 4 – LHP – C is chosen because although it does not have the lowest MAE, it demonstrates good spatial UHI bias correction, which is an important requirement for urban heatwave T prediction. With increasing number of sites, all ML models continue to show a decreasing tendency in MAE, even as the number of sites included is increased up to the limit of 92. This suggests further improvements in results could be obtained with a larger number of available observation sites. The WOW site density is approximately 0.01 km^{-2} ($=115/(110\times90)$). ML model improvements might be made by using denser CWS datasets, for example, Netatmo, which has site density of approximately 0.85 and 0.86 km^{-2} for Amsterdam and Toulouse, respectively (Fenner et al., 2021). Such site densities are approaching LCZ and neighbourhood resolving scales, where better

information on the relationship between land cover and T is available, and observation QC could be improved through nearest neighbourhood comparisons.

The influence of the number of data points (Figure 10b) and heatwaves used in training (Figure 10c) is also investigated. Up to 99.9% of data points are randomly dropped from the seven training heatwaves. When very few data points ($\sim 1\%$) are used, increasing the number of data points improves the MAE for all models. For XGB and MLP, a minimum in MAE occurs at $\approx 10\%$, and further increases in the number of data points slightly degrade MAE, which suggests overfitting occurs. However, increasing the number of heatwaves in training from 1 to 7 improves MAE for all ML models, since increasing the number of heatwaves in training means the ML models generalize better to other heatwaves. This offers an explanation for the overfitting that occurs for XGB and MLP when an increasing percentage of data points are included beyond $\approx 10\%$. The more data points available, the more the ML models can tune to the training heatwaves, and the poorer they generalize. For future improvements in ML model performance, increasing the number of heatwaves in training is more important than increasing the heatwave time sampling. This is because sampling more modes of variability in the UKV T bias is more important than having higher frequency sampling of the modes currently seen in training.

3.6 | CWS uncertainty and implications for ML

To investigate the influence of CWS uncertainty on ML model performance, each ML model is tested using data from five professionally maintained (MO) sites. Configuration $3-\mathrm{SHP}-\mathrm{C}^b$ is used since it includes built-up fraction predictors and also had reasonable performance across ML models (RFR, XGB, MLP). The ML models and UKV are more accurate at the MO sites than the WOW sites (see Figure 6, right y-axes). This might be explained by the fact that the MO observations are included in the UKV data assimilation (unlike the WOW observations) so that the initial conditions are more accurate at these sites. Furthermore, the UKV has been evaluated at the MO sites previously and has been developed to give good predictions there. Other possible explanations are that the WOW observations have a more complex siting (e.g., being close to buildings and trees) and that there is larger uncertainty in the WOW observation quality.

ML models yield improved MAE at MO sites compared to the UKV with improvements of 0.06, 0.03 and 0.02°C for RFR, XGB and MLP, respectively. However, improvements are smaller when testing at MO sites compared to when testing at the WOW sites (see Figure 5) with 0.05, 0.08 and 0.10°C lower improvements for RFR, XGB and MLP, respectively.

Compared to the observed MO composite diurnal T profile, the RFR, XGB, MLP and UKV profiles have MAEp of 0.40, 0.44, 0.41 and 0.37°C (Figure 6b), respectively. Therefore, the UKV has slightly smaller MAEp compared to the ML models, unlike when predictions are made at the WOW sites (see Section 3.3). The ML models are closer to the MO composite diurnal profile than the UKV between late evening (20:00 UTC) and early morning (08:00 UTC), but during the day there is a $\sim 1^{\circ}$C warm bias in the ML models. The reason that the ML models can have degraded MAEp while having improved MAE compared to the UKV at MO sites is that the ML models better represent T spatial variability. Note the MAEp calculation involves calculating the MAE after compositing over sites (i.e., space).

It is perhaps surprising that the ML model and UKV predictions are so similar for the WOW sites compared to the MO sites (Figure 6a,b, respectively), when one considers that the observed WOW and MO composite diurnal profiles are quite different, with the MO observed profile being ~ 1°C cooler than the WOW observed profile during the day. There are several possible explanations. The 100 m grid cell site average land covers are 0.42 built up, 0.40 tree cover, 0.16 grassland for WOW and 0.26 built up, 0.17 tree cover and 0.54 grassland for MO, so the MO sites are generally less built up. Therefore, it is possible the ML models should be cooler at the less built-up MO sites during the day, and that they do not learn that relationship. However, the urban influence on T tends to be greatest at night, so if this were the case one would expect the MO observations to be cooler during the night as well as the day. The WOW observations are often located in gardens and near building walls, so it is possible that there are micro-scale causes of high daytime T. The WOW sites could be warmer than the MO sites due to a bias associated with insufficient radiation shielding and/or ventilation, as found by other investigations of CWS data (Bell et al., 2015; Fenner et al., 2017; Meier et al., 2017). This would result in the ML models learning the bias from the WOW sites and consequently on average overestimating T at the MO sites. This might also partly explain why the ML model MAE improvements are smaller at the MO sites than at the WOW sites, and why the UKV and ML model MAEp are ~ 1°C larger in the late morning and afternoon compared to the rest of the day at the WOW sites (Figure 6a). In essence, if the WOW observations have insufficient radiation shielding and/or ventilation, then during the middle of the day the ML bias

corrections are moving the UKV towards observations that are too warm. Further analysis of WOW radiation bias and correction methods (e.g., building on the work of Cornes et al. (2020)) will be the topic of future work. This will be important for developing both predictive capability and trust in post-processing methods that utilize CWS observations for bias correcting and downscaling NWP and GCM output.

4 | CONCLUSIONS

4.1 | Summary

A ML method has been developed that has the capability to bias correct and downscale operational NWP T forecasts from 1.5 km to 100 m horizontal grid length, using WOW observations and high-resolution land cover. The proof of concept study focuses on eight heatwave cases (2019–2021) over London, UK. The performance of three ML algorithms (RFR, XGB and MLP) at predicting T and the UHI both temporally and spatially is evaluated.

The best performing ML models for RFR, XGB and MLP algorithms all give T MAE improvements of 0.12°C (11%) over the UKV, which already has a state of the art urban surface representation for operational NWP. For the special case of testing 'in sample' (i.e., where predictions are evaluated at sites included in ML model training), the best performing ML models for RFR, XGB and MLP give improvements of 0.27, 0.26 and 0.19°C, respectively. This demonstrates that that the ML method can also be used to improve NWP T predictions at specific locations where observations are available, in addition to making predictions that generalize well to locations unseen in training (i.e., when making spatially continuous T maps).

The UHI MAEp for RFR, XGB and MLP is 0.19, 0.15 and 0.20°C, respectively, which is much reduced compared to the UKV that has a MAEp of 0.64°C. The reduction in MAEp is achieved by lowering the overestimation of UKV T at urban relative to vegetated sites in the afternoon and evening. The ability of the ML method to bias correct the city-scale spatial representation of the UHI is demonstrated with T maps, where, for example, in the evening, central London is made cooler, but the more vegetated suburbs and rural surroundings are made warmer by the ML. RFR feature importance shows latent heat flux to be by far the most important predictor. The ML method is able to downscale T with qualitatively expected behaviours. For example, vegetated areas such as parks become cooler relative to more dense urban areas, and downstream regions of parks are cooler than upstream regions, via modelling the effects of upstream built-up fraction.

Compared with the ML models, a simple statistical model (MLR) performed almost as well for T MAE and composite diurnal profile prediction, but could not match the performance of ML models in bias correcting the UHI. This is consistent with linear models not being able to capture the complex relationships required to accurately bias correct the UHI.

There is a trade-off between using ML models with large and small hyperparameters for UHI and T prediction. Biggest improvements in the UHI representation are made with large hyperparameters and when built-up fraction (at the site and upstream of the site) predictors are included in addition to the UKV predictors. The former is likely because large decision trees (RFR, XGB) and neural networks (MLP) are required to learn the complex diurnally varying relationships between predictors that control the UHI. However, compared to the ML models with small hyperparameters, those with large hyperparameters have poorer T MAEs. This is particularly the case when land cover (e.g., built-up fraction) predictors are included. This is due to overfitting.

4.2 | Discussion

Although this study is limited to Greater London and eight heatwaves, the ML method could be used to incorporate data from WOW sites across the UK and from the entire WOW record period and bias correct and downscale the UKV over its entire domain. In fact, we found that increasing the number of training WOW sites and heatwaves results in T MAE still decreasing at the maximum available number of sites and heatwaves. It is therefore possible that by including WOW observations from other locations across the UK and including longer observations periods, that the ML models will not only be able to used outside of the current study region and observations periods but improve ML model predictions inside the current study region and observations periods. Increasing the density of observations could also be investigated by including Netatmo observations which are available for 2020 (Netatmo, 2021). Also, including observations from as many spatial locations and weather systems as possible should help combat overfitting, enabling more complex ML model architectures to be used. In addition, when extending the study region, other ML model predictors should be considered for inclusion, for example, building material and surface roughness properties (Brousse et al., 2023; Wang et al., 2023), orography and sea surface temperature. Following the recommendations of Wang et al. (2023), time lagged predictors could also be investigated to obtain improved temporal predictions.

An important next step towards CWS observation-based ML post-processing techniques in operational

NWP post-processing workflows is to demonstrate that ML outperforms state of the art conventional techniques (e.g., the UKMO's IMPROVER; Roberts et al., 2023), both at professional and CWS sites. This raises the question of whether CWS (in our case WOW) observations can be treated as 'truth'. In the present study, it is likely that there are WOW radiation bias issues (see Section 3.6) consistent with other CWS studies (Bell et al., 2015; Fenner et al., 2021). Development of the QC method is required to address this (e.g., following Cornes et al. (2020)). The uncertainty of QCd observations should be estimated using dense professional observations from urban field campaigns. It is suggested that a criterion for CWS to be used as 'truth' in evaluating model predictions is that the typical error of conventional post-processed NWP predictions at professional sites should be larger than the observation uncertainty at CWS sites (i.e., model error dominates observation error).

Our bias correction and downscaling method have the potential to remove the need for hectometric NWP in making accurate hectometric T predictions. This is because near-surface variables are strongly forced by the surface and may not need hectometric representation of the entire atmospheric boundary layer (and above) to be accurately predicted. Whether our ML method for bias correcting and downscaling kilometre scale NWP T has comparable skill compared to hectometric conventional NWP T should be investigated in a 'like for like' comparison, in particular using the same land cover. This should be done at forecast lead times ranging from hours to several days since we demonstrate that ML post-processed T improves relative to the UKV T with increasing forecast lead time.

AUTHOR CONTRIBUTIONS

Lewis P. Blunn: Conceptualization (lead); data curation (lead); formal analysis (lead); investigation (lead); methodology (lead); project administration (lead); supervision (lead); writing – original draft (lead); writing – review and editing (lead). **Flynn Ames:** Data curation (supporting); formal analysis (supporting); investigation (supporting); methodology (supporting); writing – original draft (supporting); writing – review and editing (supporting). **Hannah L. Croad:** Data curation (supporting); formal analysis (supporting); investigation (supporting); methodology (supporting); writing – original draft (supporting); writing – review and editing (supporting). **Adam Gainford:** Data curation (supporting); formal analysis (supporting); investigation (supporting); methodology (supporting); writing – original draft (supporting); writing – review and editing (supporting). **Ieuan Higgs:** Data curation (supporting); formal analysis (supporting); investigation (supporting); methodology (supporting); writing – original draft (supporting); writing – review and editing (supporting). **Mathew Lipson:** Data curation (supporting); writing – original draft (supporting); writing – review and editing (supporting). **Chun Hay Brian Lo:** Data curation (supporting); formal analysis (supporting); investigation (supporting); methodology (supporting); writing – original draft (supporting); writing – review and editing (supporting).

ACKNOWLEDGEMENTS

LPB was funded by the Met Office Weather and Climate Science for Service Partnership (WCSSP) India project which is supported by the Department for Science, Innovation & Technology (DSIT). HLC, AG, and BL were funded by the SCENARIO NERC Doctoral Training Partnership grant NE/S007261/1. FA was funded by a Science and Technology Facilities Council (STFC) studentship (ST/W507763/1). IH acknowledges the support of the Natural Environment Research Council via the National Centre for Earth Observation (Contract Number PR140015). The authors would like to thank Thorwald Stein and Humphrey Lean for helping coordinate the study.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Lewis P. Blunn https://orcid.org/0000-0002-3207-5002
Flynn Ames https://orcid.org/0000-0003-4915-2163
Hannah L. Croad https://orcid.org/0000-0002-5124-4860
Adam Gainford https://orcid.org/0000-0003-2484-8316
Ieuan Higgs https://orcid.org/0000-0002-3525-4962
Mathew Lipson https://orcid.org/0000-0001-5322-1796
Chun Hay Brian Lo https://orcid.org/0000-0001-7661-7080

REFERENCES

Alonso, L. & Renard, F. (2020) A new approach for understanding urban microclimate by integrating complementary predictors at different scales in regression and machine learning models. *Remote Sensing*, 12, 2434.

Barlow, J.F. (2014) Progress in observing and modelling the urban boundary layer. *Urban Climate*, 10, 216–240.

Bell, S., Cornford, D. & Bastin, L. (2015) How good are citizen weather stations? Addressing a biased opinion. *Weather*, 70, 75–84.

Best, M., Pryor, M., Clark, D., Rooney, G., Essery, R., Ménard, C. et al. (2011) The joint UK land environment simulator (JULES), model description–part 1: energy and water fluxes. *Geoscientific Model Development*, 4, 677–699.

Bohnenstengel, S.I., Evans, S., Clark, P.A. & Belcher, S.E. (2011) Simulations of the London urban heat Island. *Quarterly Journal of the Royal Meteorological Society*, 137, 1625–1640.

Breiman, L. (2001) Random forests. *Machine Learning*, 45, 5–32.

Brousse, O., Simpson, C., Kenway, O., Martilli, A., Krayenhoff, E.S., Zonato, A. et al. (2023) Spatially explicit correction of simulated urban air temperatures using crowdsourced data. *Journal of Applied Meteorology and Climatology*, 62, 1539–1572.

Brousse, O., Simpson, C., Walker, N., Fenner, D., Meier, F., Taylor, J. et al. (2022) Evidence of horizontal urban heat advection in london using six years of data from a citizen weather station network. *Environmental Research Letters*, 17, 044041.

Bunce, R., Barr, C., Clarke, R., Howard, D. & Lane, A. (1990) ITE land classification of Great Britain 1990. https://doi.org/10.5285/ab320e08-faf5-48e1-9ec9-77a213d2907f

Chantry, M., Christensen, H., Dueben, P. & Palmer, T. (2021) Opportunities and challenges for machine learning in weather and climate modelling: hard, medium and soft AI. *Philosophical Transactions of the Royal Society A*, 379, 20200083.

Chase, R.J., Harrison, D.R., Burke, A., Lackmann, G.M. & McGovern, A. (2022) A machine learning tutorial for operational meteorology. Part I: traditional machine learning. *Weather and Forecasting*, 37, 1509–1529.

Chen, G., Hua, J., Shi, Y. & Ren, C. (2023) Constructing air temperature and relative humidity-based hourly thermal comfort dataset for a high-density city using machine learning. *Urban Climate*, 47, 101400.

Chen, S., Yang, Y., Deng, F., Zhang, Y., Liu, D., Liu, C. et al. (2022) A high-resolution monitoring approach of canopy urban heat Island using a random forest model and multi-platform observations. *Atmospheric Measurement Techniques*, 15, 735–756.

Cho, D., Yoo, C., Im, J. & Cha, D.-H. (2020) Comparative assessment of various machine learning-based bias correction methods for numerical weather prediction model forecasts of extreme air temperatures in urban areas. *Earth and Space Science*, 7, e2019EA000740.

Clark, D., Mercado, L., Sitch, S., Jones, C., Gedney, N., Best, M. et al. (2011) The joint UK land environment simulator (JULES), model description–part 2: carbon fluxes and vegetation dynamics. *Geoscientific Model Development*, 4, 701–722.

Clark, M.R., Webb, J.D. & Kirk, P.J. (2018) Fine-scale analysis of a severe hailstorm using crowd-sourced and conventional observations. *Meteorological Applications*, 25, 472–492.

Cornes, R.C., Dirksen, M. & Sluiter, R. (2020) Correcting citizen-science air temperature measurements across The Netherlands for short wave radiation bias. *Meteorological Applications*, 27, e1814.

Davies, T., Cullen, M.J., Malcolm, A.J., Mawson, M., Staniforth, A., White, A. et al. (2005) A new dynamical core for the met Office's global and regional modelling of the atmosphere. *Quarterly Journal of the Royal Meteorological Society*, 131, 1759–1782.

Demuzere, M., Kittner, J., Martilli, A., Mills, G., Moede, C., Stewart, I.D. et al. (2022) A global map of local climate zones to support earth system modelling and urban scale environmental science. *Earth System Science Data Discussions*, 2022, 1–57.

Demuzere, M., Kittner, J., Martilli, A., Mills, G., Moede, C., Stewart, I.D. et al. (2024) Google earth engine: global map of local climate zones. URL https://developers.google.com/earth-engine/datasets/catalog/RUB_RUBCLIM_LCZ_global_lcz_map_latest

dos Santos, R.S. (2020) Estimating spatio-temporal air temperature in london (UK) using machine learning and earth observation satellite data. *International Journal of Applied Earth Observation and Geoinformation*, 88, 102066.

Droste, A., Pape, J.-J., Overeem, A., Leijnse, H., Steeneveld, G.-J., Van Delden, A. et al. (2017) Crowdsourcing urban air temperatures through smartphone battery temperatures in São Paulo, Brazil. *Journal of Atmospheric and Oceanic Technology*, 34, 1853–1866.

DUKES. (2003) Digest of United Kingdom energy statistics 2003. URL https://webarchive.nationalarchives.gov.uk/ukgwa/20031221111208/http://www.dti.gov.uk/energy/inform/dukes/dukes2003/index.shtml

Espeholt, L., Agrawal, S., Sønderby, C., Kumar, M., Heek, J., Bromberg, C. et al. (2022) Deep learning for twelve hour precipitation forecasts. *Nature Communications*, 13, 5145.

Feichtinger, M., de Wit, R., Goldenits, G., Kolejka, T., Hollósi, B., Žuvela-Aloise, M. et al. (2020) Case-study of neighborhood-scale summertime urban air temperature for the City of Vienna using crowd-sourced data. *Urban Climate*, 32, 100597.

Fenner, D., Bechtel, B., Demuzere, M., Kittner, J. & Meier, F. (2021) Crowdqc+—a quality-control for crowdsourced air-temperature observations enabling world-wide urban climate applications. *Frontiers in Environmental Science*, 9, 553.

Fenner, D., Meier, F., Bechtel, B., Otto, M. & Scherer, D. (2017) Intra and inter local climate zone variability of air temperature as observed by crowdsourced citizen weather stations in Berlin, Germany. *Meteorologische Zeitschrift*, 26, 525–547.

Friedman, J.H. (2001) Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29, 1189–1232.

Garcia-Marti, I., Overeem, A., Noteboom, J.W., de Vos, L., de Haij, M. & Whan, K. (2023) From proof-of-concept to proof-of-value: approaching third-party data to operational workflows of national meteorological services. *International Journal of Climatology*, 43, 275–292.

Gardner, M.W. & Dorling, S. (1998) Artificial neural networks (the multilayer perceptron) – a review of applications in the atmospheric sciences. *Atmospheric Environment*, 32, 2627–2636.

Gettelman, A., Gagne, D.J., Chen, C.-C., Christensen, M., Lebo, Z., Morrison, H. et al. (2021) Machine learning the warm rain process. *Journal of Advances in Modeling Earth Systems*, 13, e2020MS002268.

Green, H.K., Andrews, N., Armstrong, B., Bickler, G. & Pebody, R. (2016) Mortality during the 2013 heatwave in England–how did it compare to previous heatwaves? A retrospective observational study. *Environmental Research*, 147, 343–349.

Grimmond, C.S.B., Blackett, M., Best, M.J., Baik, J.-J., Belcher, S., Beringer, J. et al. (2011) Initial results from phase 2 of the international urban energy balance model comparison. *International Journal of Climatology*, 31, 244–272.

Grimmond, C.S.B., Blackett, M., Best, M.J., Barlow, J., Baik, J., Belcher, S. et al. (2010) The international urban energy balance models comparison project: first results from phase 1. *Journal of Applied Meteorology and Climatology*, 49, 1268–1292.

Hahn, C., Garcia-Marti, I., Sugier, J., Emsley, F., Beaulant, A.-L., Oram, L. et al. (2022) Observations from personal weather stations—eumetnet interests and experience. *Climate*, 10, 192.

Ham, Y.-G., Kim, J.-H. & Luo, J.-J. (2019) Deep learning for multi-year enso forecasts. *Nature*, 573, 568–572.

Hammerberg, K., Brousse, O., Martilli, A. & Mahdavi, A. (2018) Implications of employing detailed urban canopy parameters for mesoscale climate modelling: a comparison between wudapt and gis databases over vienna, Austria. *International Journal of Climatology*, 38, e1241–e1257.

Harris, L., McRae, A.T., Chantry, M., Dueben, P.D. & Palmer, T.N. (2022) A generative deep learning approach to stochastic downscaling of precipitation forecasts. *Journal of Advances in Modeling Earth Systems*, 14, e2022MS003120.

Ho, T.K. (1995) Random decision forests. In: *Proceedings of 3rd international conference on document analysis and recognition*, Vol. 1, pp. 278–282. IEEE. https://scholar.google.co.uk/scholar?hl=en&as_sdt=0%2C5&q=Ho%2C+T.K.+%281995%29+Random+decision+forests.+In%3A+Proceedings+of+3rd+international+conference+on+document+analysis+and+recognition%2C+Vol.+1%2C+pp.+278%E2%80%93282.&btnG=

Kaae Sønderby, C., Espeholt, L., Heek, J., Dehghani, M., Oliver, A., Salimans, T. et al. (2020) Metnet: a neural weather model for precipitation forecasting. *arXiv e-prints*, arXiv–2003.

Kashinath, K., Mustafa, M., Albert, A., Wu, J., Jiang, C., Esmaeilzadeh, S. et al. (2021) Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A*, 379, 20200093.

Keisler, R. (2022) Forecasting global weather with graph neural networks. arXiv Preprint arXiv:2202.07575.

Keras Developers. (2023) Drop out layers. URL https://keras.io/api/layers/regularization_layers/dropout/

Kirk, P.J., Clark, M.R. & Creed, E. (2021) Weather observations website. *Weather*, 76, 47–49.

Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Pritzel, A. et al. (2022) GraphCast: learning skillful medium-range global weather forecasting. arXiv Preprint arXiv:2212.12794.

Lipson, M.J., Grimmond, S., Best, M., Abramowitz, G., Coutts, A., Tapper, N. et al. (2023) Evaluation of 30 urban land surface models in the urban-plumber project: phase 1 results. *Quarterly Journal of the Royal Meteorological Society*, 150, 126–169.

Lopez-Gomez, I., McGovern, A., Agrawal, S. & Hickey, J. (2023) Global extreme heat forecasting using neural weather models. *Artificial Intelligence for the Earth Systems*, 2, e220035.

Lyu, F., Wang, S., Han, S.Y., Catlett, C. & Wang, S. (2022) An integrated cyberGIS and machine learning framework for fine-scale prediction of urban Heat Island using satellite remote sensing and urban sensor network data. *Urban Informatics*, 1, 6.

Masson, V., Heldens, W., Bocher, E., Bonhomme, M., Bucher, B., Burmeister, C. et al. (2020) City-descriptive input data for urban climate models: model requirements, data sources and challenges. *Urban Climate*, 31, 100536.

Meier, F., Fenner, D., Grassmann, T., Otto, M. & Scherer, D. (2017) Crowdsourcing air temperature from citizen weather stations for urban climate research. *Urban Climate*, 19, 170–191.

Met Office (2022) MIDAS open: UK hourly weather observation data, v202207. NERC EDS Centre for Environmental Data Analysis. https://doi.org/10.5285/6180fb7ed76a442eb1b8f3f152fd08d7.

Met Office. (2024) Weather observations website. URL https://wow.metoffice.gov.uk/sites/search

Met Office Hadley Centre. (2024) Centre Hadley Central England temperature (HadCET) dataset. URL http://www.metoffice.gov.uk/hadobs/hadcet/index.html

Meyer, D., Grimmond, S., Dueben, P., Hogan, R. & van Reeuwijk, M. (2022) Machine learning emulation of urban land surface processes. *Journal of Advances in Modeling Earth Systems*, 14, e2021MS002744.

Meyer, D., Hogan, R.J., Dueben, P.D. & Mason, S.L. (2022) Machine learning emulation of 3d cloud radiative effects. *Journal of Advances in Modeling Earth Systems*, 14, e2021MS002550.

Milan, M., Macpherson, B., Tubbs, R., Dow, G., Inverarity, G., Mittermaier, M. et al. (2020) Hourly 4d-var in the met office ukv operational forecast model. *Quarterly Journal of the Royal Meteorological Society*, 146, 1281–1301.

Milojevic-Dupont, N. & Creutzig, F. (2021) Machine learning for geographically differentiated climate change mitigation in urban areas. *Sustainable Cities and Society*, 64, 102526.

Mitchell, T.D. & Fry, M.J. (2024) The importance of crowdsourced observations for urban climate services. *International Journal of Climatology*, 44, 1409–1422.

Muller, C.L., Chapman, L., Grimmond, C., Young, D.T. & Cai, X. (2013) Sensors and the city: a review of urban meteorological networks. *International Journal of Climatology*, 33, 1585–1600.

Nazarian, N., Krayenhoff, E., Bechtel, B., Hondula, D., Paolini, R., Vanos, J. et al. (2022) Integrated assessment of urban overheating impacts on human life. *Earth's Future*, 10, e2022EF002682.

Nazarian, N., Liu, S., Kohler, M., Lee, J.K., Miller, C., Chow, W.T. et al. (2021) Project Coolbit: can your watch predict heat stress and thermal comfort sensation? *Environmental Research Letters*, 16, 034031.

Netatmo. (2021) *EUMETNET sandbox: Netatmo observing network data v1*. NERC EDS Centre for Environmental Data Analysis. URL. Available from: https://catalogue.ceda.ac.uk/uuid/e8793d74a651426692faa100e3b2acd3

Netatmo. (2023) Netatmo personal weather station. URL https://www.netatmo.com/en-us/

Oke, T.R., Mills, G., Christen, A. & Voogt, J.A. (2017) *Urban climates*. Cambridge University Press. https://scholar.google.co.uk/scholar?hl=en&as_sdt=0%2C5&q=oke+urban+climate&btnG=#d=gs_cit&t=1715161771578&u=%2Fscholar%3Fq%3Dinfo%3AFGFxx9Ou3ZAJ%3Ascholar.google.com%2F%26output%3Dcite%26scirp%3D0%26hl%3Den

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M. et al. (2022) Fourcastnet: a global data-driven high-resolution weather model using adaptive fourier neural operators. arXiv Preprint arXiv:2202.11214.

PHE. (2019) PHE heatwave mortality monitoring: summer 2019. URL https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/942646/PHE_heatwave_report_2019.pdf

PHE. (2020) Heatwave mortality monitoring report: 2020. URL https://www.gov.uk/government/publications/phe-heatwave-mortality-monitoring/heatwave-mortality-monitoring-report-2020

PHE. (2021) Heatwave mortality monitoring report: 2021. URL https://www.gov.uk/government/publications/heat-mortality-monitoring-reports/heat-mortality-monitoring-report-2021

Porson, A., Clark, P.A., Harman, I., Best, M. & Belcher, S. (2010) Implementation of a new urban energy budget scheme in the MetUM. Part I: description and idealized simulations. *Quarterly Journal of the Royal Meteorological Society*, 136, 1514–1529.

Potgieter, J., Nazarian, N., Lipson, M.J., Hart, M.A., Ulpiani, G., Morrison, W. et al. (2021) Combining high-resolution land use data with crowdsourced air temperature to investigate intra-urban microclimate. *Frontiers in Environmental Science*, 9, 385.

Rasp, S., Pritchard, M.S. & Gentine, P. (2018) Deep learning to represent subgrid processes in climate models. *Proceedings of the National Academy of Sciences*, 115, 9684–9689.

Rasp, S. & Thuerey, N. (2021) Data-driven medium-range weather prediction with a resnet pretrained on climate simulations: a new model for weatherbench. *Journal of Advances in Modeling Earth Systems*, 13, e2020MS002405.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P. et al. (2021) Skilful precipitation nowcasting using deep generative models of radar. *Nature*, 597, 672–677.

Rawlins, F., Ballard, S., Bovis, K., Clayton, A., Li, D., Inverarity, G. et al. (2007) The met office global four-dimensional variational data assimilation scheme. *Quarterly Journal of the Royal Meteorological Society: A Journal of the Atmospheric Sciences, Applied Meteorology and Physical Oceanography*, 133, 347–362.

Roberts, N., Ayliffe, B., Evans, G., Moseley, S., Rust, F., Sandford, C. et al. (2023) IMPROVER: the new probabilistic postprocessing system at the met Office. *Bulletin of the American Meteorological Society*, 104, E680–E697.

Rolnick, D., Donti, P.L., Kaack, L.H., Kochanski, K., Lacoste, A., Sankaran, K. et al. (2022) Tackling climate change with machine learning. *ACM Computing Surveys (CSUR)*, 55, 1–96.

Ronda, R., Steeneveld, G., Heusinkveld, B., Attema, J. & Holtslag, A. (2017) Urban finescale forecasting reveals weather conditions with unprecedented detail. *Bulletin of the American Meteorological Society*, 98, 2675–2688.

Schoetter, R., Kwok, Y.T., de Munck, C., Lau, K.K.L., Wong, W.K. & Masson, V. (2020) Multi-layer coupling between SURFEX-TEB-v9. 0 and Meso-NH-v5. 3 for modelling the urban climate of high-rise cities. *Geoscientific Model Development*, 13, 5609–5643.

scikit-learn Developers. (2023a) LinearRegression. URL https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

scikit-learn Developers. (2023b) RandomForestRegressor. URL https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

scikit-learn Developers. (2023c) sklearn Version 1.1.3. URL https://scikit-learn.org/1.1/

Shi, X., Gao, Z., Lausen, L., Wang, H., Yeung, D.-Y., Wong, W.-K. et al. (2017) Deep learning for precipitation nowcasting: a benchmark and a new model. *Advances in Neural Information Processing Systems*, 30. https://scholar.google.co.uk/scholar?hl=en&as_sdt=0%2C5&q=Deep+learning+for+precipitation+nowcasting%3A+a+benchmark+and+a+new+model&btnG=#d=gs_cit&t=1715161989389&u=%2Fscholar%3Fq%3Dinfo%3AAVXu3ucg-Z4J%3Ascholar.google.com%2F%26output%3Dcite%26scirp%3D0%26hl%3Den

Skamarock, W., Klemp, J., Dudhia, J., Gill, D., Liu, Z., Berner, J. and Huang, X. (2018) A description of the advanced research wrf model version 4.3 (july). National center for atmospheric research. URL: https://doi.org/10.5065/1dfh-6p97.

Stengel, K., Glaws, A., Hettinger, D. & King, R.N. (2020) Adversarial super-resolution of climatological wind and solar data. *Proceedings of the National Academy of Sciences*, 117, 16805–16815.

Stewart, I.D. & Oke, T.R. (2012) Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society*, 93, 1879–1900.

Straub, A., Berger, K., Breitner, S., Cyrys, J., Geruschkat, U., Jacobeit, J. et al. (2019) Statistical modelling of spatial patterns of the urban heat Island intensity in the urban environment of augsburg, Germany. *Urban Climate*, 29, 100491.

Tang, Y., Lean, H.W. & Bornemann, J. (2013) The benefits of the met Office variable resolution NWP model for forecasting convection. *Meteorological Applications*, 20, 417–426.

TensorFlow Developers. (2023a) Dense. URL https://www.tensorflow.org/api_docs/python/tf/keras/layers/Dense

TensorFlow Developers. (2023b) TensorFlow Version 2.9.1. URL https://www.tensorflow.org/versions/r2.9/api_docs/python/tf

van Beekvelt, D., Garcia-Marti, I. & de Baar, J. (2024) Towards high-resolution gridded climatology stemming from the combination of official and crowdsourced weather observations using multi-fidelity methods. *PLOS Climate*, 3, e0000216.

Venter, Z.S., Brousse, O., Esau, I. & Meier, F. (2020) Hyperlocal mapping of urban air temperature using remote sensing and crowdsourced weather data. *Remote Sensing of Environment*, 242, 111791.

Vulova, S., Meier, F., Fenner, D., Nouri, H. & Kleinschmit, B. (2020) Summer nights in Berlin, Germany: modeling air temperature spatially with remote sensing, crowdsourced weather data, and machine learning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13, 5074–5087.

Wang, H., Yang, J., Chen, G., Ren, C. & Zhang, J. (2023) Machine learning applications on air temperature prediction in the urban canopy layer: a critical review of 2011–2022. *Urban Climate*, 49, 101499.

Weyn, J.A., Durran, D.R., Caruana, R. & Cresswell-Clay, N. (2021) Sub-seasonal forecasting with a large ensemble of deep-learning weather prediction models. *Journal of Advances in Modeling Earth Systems*, 13, e2021MS002502.

WMO. (2018) Guide to instruments and methods of observation. URL https://library.wmo.int/doc_num.php?explnum_id=11386s

Wood, N., Staniforth, A., White, A., Allen, T., Diamantakis, M., Gross, M. et al. (2014) An inherently mass-conserving semi-implicit semi-Lagrangian discretization of the deep-atmosphere global non-hydrostatic equations. *Quarterly Journal of the Royal Meteorological Society*, 140, 1505–1520.

Wu, Y., Teufel, B., Sushama, L., Belair, S. & Sun, L. (2021) Deep learning-based super-resolution climate simulator-emulator

framework for urban heat studies. *Geophysical Research Letters*, 48, e2021GL094737.

XGBoost Developers. (2023a) xgboost Version 1.7.1. URL https://pypi.org/project/xgboost/1.7.1/

XGBoost Developers. (2023b) XGBRegressor. URL https://xgboost.readthedocs.io/en/latest/python/python_api.html

Yu, Z., Chen, S., Wong, N.H., Ignatius, M., Deng, J., He, Y. et al. (2020) Dependence between urban morphology and outdoor air temperature: a tropical campus study using random forests algorithm. *Sustainable Cities and Society*, 61, 102200.

Zanaga, D., van de Kerchove, R., de Keersmaecker, W., Souverijns, N., Brockmann, C., Quast, R. et al. (2021) Esa worldcover 10 m 2020 v100. 2021.

Zumwald, M., Knüsel, B., Bresch, D.N. & Knutti, R. (2021) Mapping urban temperature using crowd-sensing data and machine learning. *Urban Climate*, 35, 100739.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Blunn, L. P., Ames, F., Croad, H. L., Gainford, A., Higgs, I., Lipson, M., & Lo, C. H. B. (2024). Machine learning bias correction and downscaling of urban heatwave temperature predictions from kilometre to hectometre scale. *Meteorological Applications*, *31*(3), e2200. https://doi.org/10.1002/met.2200

APPENDIX A: MACHINE LEARNING MODEL CONFIGURATIONS TABLE

TABLE A1 Predictor and hyperparameter configurations.

Configuration name	Predictors	RFR (g, h, k)	XGB (g, k, η, γ)	MLP ($l-m-n$, p, r)
1 – SHP – C	T	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
1 – SHP – K^*	$T + K^*$	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
1 – SHP – CL	$T + CL$	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
1 – SHP – WS	$T + WS$	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
1 – SHP – RH	$T + RH$	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
1 – SHP – Q_E	$T + Q_E$	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
1 – SHP – Q_H	$T + Q_H$	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
1 – SHP – Q_{soil}	$T + Q_{soil}$	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
1 – SHP – SM	$T + SM$	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
1 – SHP – HoD	$T + HoD$	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
2 – SHP – C	UKV	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
2 – SHP – $BUp1^b$	UKV + $BUp1^b$	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
2 – SHP – $TCp1^b$	UKV + $TCp1^b$	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
2 – SHP – $GLp1^b$	UKV + $GLp1^b$	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
2 – SHP – $PWp1^b$	UKV + $PWp1^b$	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
2 – SHP – $BU1^b$	UKV + $BU1^b$	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
2 – SHP – $BU5^b$	UKV + $BU5^b$	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
2 – SHP – $BU25^b$	UKV + $BU25^b$	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
2 – SHP – $BUp1diff^b$	UKV + $BUp1diff^b$	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
2 – SHP – $BU1diff^b$	UKV + $BU1diff^b$	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
2 – SHP – $p1^b$	UKV + $p1^b$ LC	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
2 – SHP – 1^b	UKV + 1^b LC	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
2 – SHP – 5^b	UKV + 5^b LC	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
2 – SHP – 25^b	UKV + 1^b LC	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
2 – SHP – $p1diff^b$	UKV + $p1^b$ LC diff.	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
2 – SHP – $1diff^b$	UKV + 1^b LC diff.	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
3 – SHP – C	UKV + URB LC	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
3 – SHP – C^b	UKV + URB^b LC	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
4 – LHP – C	UKV + URB LC	25, 15, 1	25, 8, 0.3, 0	32 – 32, 15, 32
4 – LHP – C^b	UKV + URB^b LC	25, 15, 1	25, 8, 0.3, 0	32 – 32, 15, 32
4 – LHP – Drop	UKV + URB LC	–	–	32 – 32, 15, 32
4 – LHP – l1l2	UKV + URB LC	–	–	32 – 32, 15, 32
5 – DHP – a	UKV + URB^b LC	25, 3, 1	25, 2, 0.3, 0	2 – 2, 5, 32
5 – DHP – b	UKV + URB^b LC	25, 5, 1	25, 5, 0.3, 0	8 – 8, 5, 32
5 – DHP – c	UKV + URB^b LC	25, 12, 1	25, 8, 0.3, 5	4 – 4 – 4, 5, 32
5 – DHP – d	UKV + URB^b LC	12, 8, 1	12, 3, 0.3, 0	4 – 4, 2, 32
5 – DHP – e	UKV + URB^b LC	100, 8, 1	100, 3, 0.3, 0	4 – 4, 15, 32
5 – DHP – f	UKV + URB^b LC	25, 8, 0.75	25, 3, 0.15, 0	4 – 4, 50, 32
6 – LHP – NoLC	UKV	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32
7 – SHP – WOWT	UKV + URB^b LC	25, 8, 1	25, 3, 0.3, 0	4 – 4, 5, 32

Note: UKV represents all UKV predictors, diff. represents the difference between ITE and World Cover land cover, URB LC represents all built-up predictors (including upstream and differences), superscript b indicates land cover is binned into 0.2 fractions, NoLC indicates no land cover was included, and WOWT indicates that the WOW T was the target. $p1$, and 1, 5 and 25 LC correspond to all land cover predictors for 100 m at the site, and 1, 5 and 25 km upstream distances, respectively. Hyperparameters g, h and k correspond to number of trees, maximum tree depth and the number of predictors considered in each tree split, respectively. Hyperparameters η and γ correspond to the learning rate and the minimum loss reduction required to make a further partition on a leaf node of the tree, respectively. Hyperparameters l, m and n are the number of nodes in the first, second and third MLP layers, respectively. Hyperparameters p and r correspond to the MLP number of epochs and batch size, respectively.