Bias correction of modelled urban temperatures with crowd-sourced weather data

Oscar Brousse, [a] Charles Simpson, [a] Owain Kenway, [b] Alberto Martilli, [c] E. Scott Krayenhoff, [d]
Andrea Zonato, [e] and Clare Heaviside, [a]

[a] *Institute of Environmental Design and Engineering, University College London*
[b] *Centre for Advanced Research Computing, University College London*
[c] *Center for Energy, Environment and Technology (CIEMAT)*
[d] *School of Environmental Sciences, University of Guelph*
[e] *Department of Civil, Environmental and Mechanical Engineering, University of Trento*

Corresponding author: O. Brousse, o.brousse@ucl.ac.uk

ABSTRACT: Urban climate model evaluation often remains limited by a lack of trusted urban weather observations. The increasing density of personal weather stations (PWS) make them a potential rich source of data for urban climate studies that address the lack of representative urban weather observations. In our study, we demonstrate that PWS data not only improve urban climate models' evaluation, but can also serve for bias-correcting their output prior to any urban climate impact studies. After simulating near-surface air temperatures over London and south-east England during the hot summer of 2018 with the Weather Research Forecast (WRF) model and its Building Effect Parameterization with the Building Energy Model (BEP-BEM) activated, we evaluated the modelled temperatures against 407 urban PWS and showcased a heterogeneous spatial distribution of the model's cool bias that was not captured using official weather stations only. This finding indicated a need for spatially-explicit urban bias corrections of air temperatures, which we performed using an innovative method using machine learning to predict the models' biases in each urban grid cell. Our technique is the first to consider that urban temperatures are heterogeneously accurate in space and that this accuracy is not linearly correlated to the urban fraction. Our results showed that the bias-correction was beneficial to bias-correct daily-minimum, -mean, and -maximum temperatures in the cities. We recommend that urban climate modellers further investigate the use of PWS for model evaluation and derive a framework for bias-correction of urban climate simulations that can serve urban climate impact studies.

SIGNIFICANCE STATEMENT: Urban climate simulations are subject to spatially heterogeneous biases in urban air temperatures. Common validation methods using official weather stations do not suffice for detecting these biases. Using a dense set of personal weather stations in London we detect these biases before proposing an innovative way for correcting them with machine learning techniques. We argue that any urban climate impact study should use such technique if possible and that urban climate scientists should continue investigating paths to improve our methods.

1. Introduction

Although decades following the 1960s have seen an increase in the body of literature on urban climates (Oke et al. 2017), the scales of applicability and the transferability of their outcomes are often limited. This can partially be attributed to the lack of observations representative of the variety of existing urban climates in cities. To address this impediment, two major solutions were proposed over the past 20 years: firstly, the development of urban surface energy balance and climate models (e.g., Masson (2000), Martilli et al. (2002), Wouters et al. (2016)) that are coupled to regional climate models, and secondly, the increased interest towards crowd-sourced and low-cost weather sensors (e.g., Muller et al. (2015), Chapman et al. (2017), Fenner et al. (2017), Meier et al. (2017)). Indeed, after proper validation and parameterization, urbanized regional climate models, with their urban climate models (UCMs) activated, offer an unprecedented opportunity to represent the impact of cities on a wide variety of weather variables at very high spatial and temporal resolutions – an opportunity further supported by the recent development of global standardized land use land cover datasets designed for urban climate studies that permit their parameterization in cities formerly deprived of these data (see the World Urban Dataset and Access Portal Tool (WUDAPT) project; Ching et al. (2018), Demuzere et al. (2022)). Likewise, after proper filtering and quality control (Napoly et al. 2018; Fenner et al. 2021), crowd-sourced personal weather sensors (PWS) permit the extension of sensing networks into urban environments that were formerly not studied despite the fact that PWS often do not meet the standards imposed by official meteorological offices for implementation of weather stations.

Nonetheless, both these tools have limitations. For instance, PWS observations are of lower reliability and accuracy than official weather stations and cover only recent years, booming after 2015 (Brousse et al. 2022). Relating the observed weather to the underlying environmental characteristics can also be difficult, as requirements for the site description are modest to non-existent. Recent efforts have however managed to bridge this information gap by using earth observations, and in particular the universal standardized Local Climate Zones land-use land-cover classification (Stewart and Oke 2012) which permitted not only the study of key weather variables in cities, like near-surface air temperature (e.g., Fenner et al. (2019), Potgieter et al. (2021), Benjamin et al. (2021), Varentsov et al. (2021)), but also their prediction via machine learning (Venter et al. 2020, 2021). Though these mapping efforts, aided by the development of machine learning algorithms, are substantial achievements, better predictions were usually achieved at low temporal resolution (e.g., weekly or seasonally), thus calling for more research to be done in order to reach improved performance at daily or hourly time-steps. This research could help comparing the outputs of physical models, like UCMs, to predicted maps of urban-specific weather variables obtained via PWS.

Indeed, UCMs are computationally expensive, require complex and energy-consuming computational infrastructures, and require prior expert-knowledge to be properly used. For example, depending on the study in terms of city-location, domain dimension, horizontal and temporal scales, or meteorological variables of interest, users need to ascertain that their models are meaningfully parameterized with the right physical and dynamical schemes, and forced by representative initial and boundary atmospheric conditions. These requirements are usually verified through sensitivity tests performed before running real-case scenarios, where model outputs are compared against weather measurements obtained from official weather stations. Then, users balance the need for accuracy with computational limitations. Notwithstanding, hazardous uncertainties exist even after sensitivity studies are performed, as shown by Bassett et al. (2020) who demonstrated that the starting time of the simulation had a noticeable impact on the modelled air temperature at 2 m in London during the summer 2018. Moreover, because of the lack of official weather stations in cities, measuring existing uncertainties per urban climate archetype is not feasible. This means that certain urban environments are poorly evaluated and hence modelled, assuming that UCMs will perform similarly under all constraints imposed by the variety of urban environments that compose a city. In face of this challenge, quality-checked crowd-sourced PWS allow monitoring for a range of urban environments. They can therefore serve the evaluation of UCMs, as Hammerberg et al. (2018) demonstrated over Vienna. But the potential of PWS may even be greater, particularly when used jointly with or in parallel to UCMs. In fact, a recent study by Sgoff et al. (2022) improved the weather forecasting of the Icosahedral Nonhydrostatic Model (ICON; Zängl et al. (2015)) at a horizontal resolution of 2 km over Germany by assimilating the data provided by PWS for air temperature and relative humidity at 2 m height. Although data assimilation is done while UCMs are running, PWS could also be used to subsequently bias-correct urban climate simulations. To date, no study has explored how PWS could be used to bias-correct simulated urban climates despite the need for realistic urban weather data of present and future urban climates to perform impact studies that can guide decision-making.

Oleson et al. (2018) already noted the need of a global dataset on urban weather observations to properly bias-correct simulated urban climates: we here propose to use the densifying network of PWS to bias-correct urban climate simulations for urban climate impact studies. Common practice in bias-correction of urban climate simulations is to apply a single correction by the mean bias at official weather stations' rural sites, thereby assuming that the urban heat island phenomenon is accurately represented by the UCM (e.g., Lauwaet et al. (2015), or Oleson et al. (2018)). Some studies however tried considering the urban effect by linearly transforming the bias-correction coefficient via an urbanization ratio calculated at each grid cell, like in Wouters et al. (2017) over Belgium. Assuming that urban climate simulations biases cannot be linearly related to the urban fraction only, we decided to test whether urban in-situ observations can be used to perform an urban-specific bias-correction of air temperatures driven by machine learning. We hereby hypothesize that such innovative bias-correction method would be beneficial for urban heat impact studies by improving the UCM outputs on which they rely. Such innovations are needed to better assess the heat burden in cities (Nazarian et al. 2022).

To respond to these questions through the scope of urban near-surface temperatures, we: i) evaluated the ability of the complex three-dimensional UCM embedded in WRF – the Building Effect Parameterization coupled with its Buidling Energy Model (BEP-BEM) – to accurately represent the urban impact on air temperatures under two boundary layer schemes for the summer 2018 in south-east England using official weather stations and PWS separately to show their added value for detecting spatially heterogeneous urban temperature biases; ii) used machine learning regressions to predict the models' daily air temperature biases in the urban environment and bias-correct the two simulations suggested in part i – which allowed us to determine an optimal time-step at which the bias-correction should be performed to optimize the outputs.; and iii) compared the two bias-corrected products against the predicted daily air temperatures using only PWS measurements to investigate how realistic the bias-corrected products are. In parallel, to illustrate the benefit gained from the bias-correction for impact studies, we showcase how the bias-correction leads to different population weighted temperatures in the Greater London area.

It is important to consider that our study does not try to estimate how a bias-corrected modelled product is better compared to a predicted product from observations for urban climate impact studies. We hereby simply try to demonstrate that any urban climate impact work that is based on urban climate modelling should pursue a spatially explicit bias-correction specific to urban areas.

2. Methods

a. Model setup and region of interest

We focused our study on the south-eastern parts of England, centred over the metropolis of London, host to approximately 9 million inhabitants. We chose to model the impact of urbanization on 2 m air temperature in London during the Summer of 2018, since it was one of the hottest summers in recent years. Indeed, the British Isles heatwaves of summer 2018 is considered to be the hottest summer on record for mean temperature (McCarthy et al. 2019), with maximum daily temperatures often over-passing 30 °C (Figure 2). Record temperatures, recently over-passed during the 2019 and 2022 summers, were recorded on the 26^{th} of July with a maximum of 34.4 °C measured at London's Heathrow airport.

To model the impact of the urban areas of London and south-east England on local meteorology, we used the Weather Research Forecast (WRF) regional climate model version 4.3 and activate the embedded Building Effect Parameterization (BEP; Martilli et al. (2002)) urban climate model with its partner Building Energy Model (BEM; Salamanca et al. (2010); Salamanca and Martilli (2010)) – hereafter referred to as BEP-BEM. We ran the model at a horizontal resolution of 1 x 1 km following a two-way nesting strategy where the outer domain is forced by ERA5 6-hourly data at 25 km with 199 by 199 grid points and the two intermediate domains are run at horizontal resolutions of 9 and 3 kilometres with 252 by 241 and 210 by 180 grid points, respectively (Figure 2, upper panel). Initial land surface conditions were provided by the default MODIS 5-arc-second land use dataset provided by the WRF community while sea surface temperatures were updated

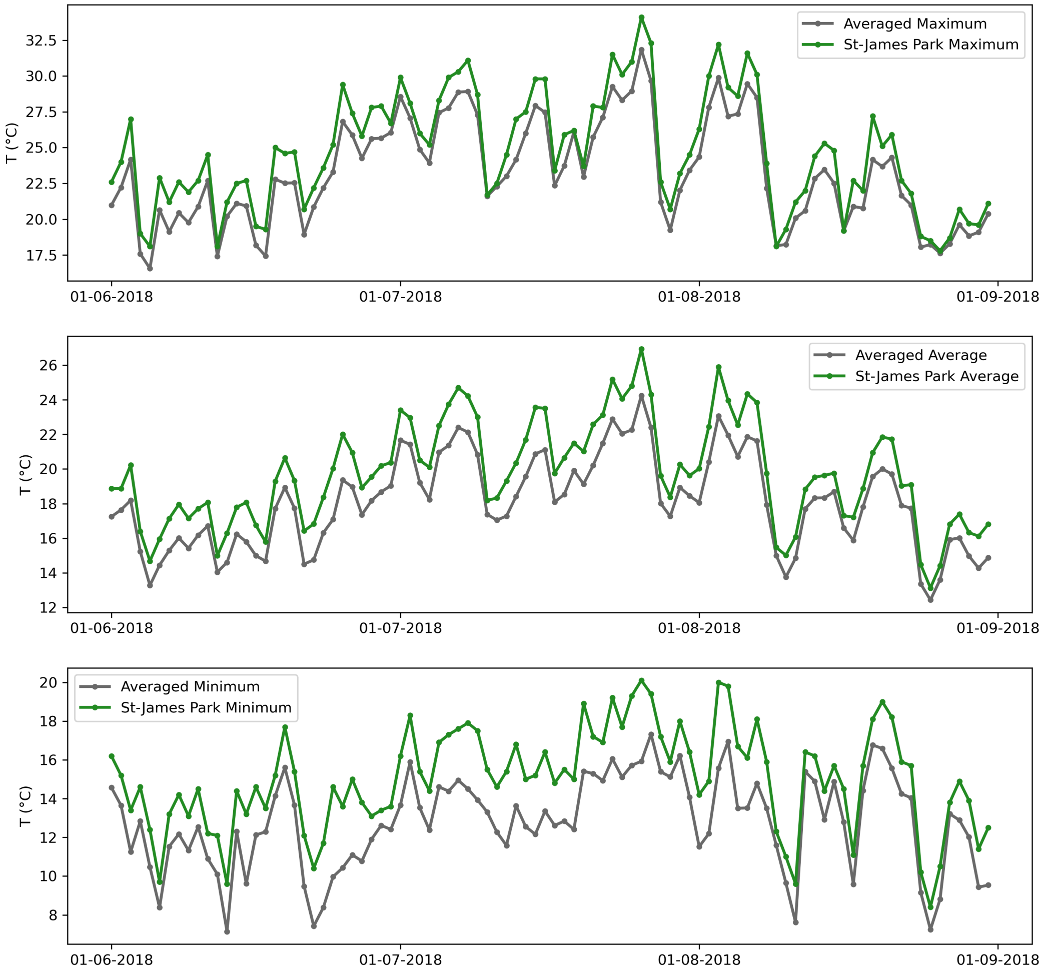


FIG. 1. Daily minimum, average and maximum temperatures observed by the Met Office MIDAS automatic weather stations. The urban St-James' Park station in central London (green) is always hotter than the average temperature of all MIDAS stations in south-east England (grey)

6-hourly out of ERA-5. We ran the model in parallel over 200 CPUs using restarts every four days of simulation. We started the simulations on the 25^{th} of May 2018 and end them on the 31^{st} of August 2018, considering the first 7 days of simulation as spin-up time.

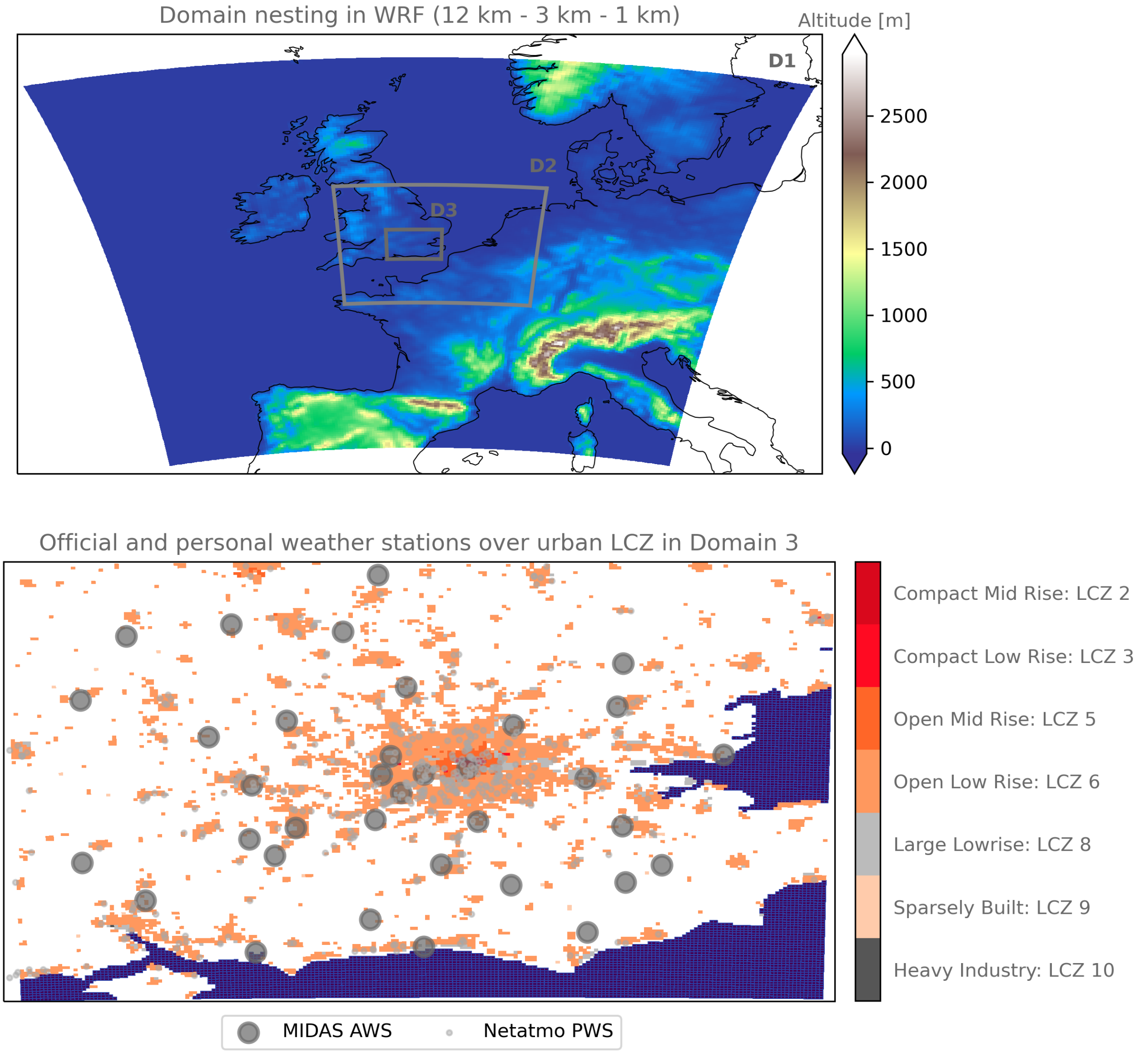
All domains used the same physical and dynamical parameterizations which we obtained out of preliminary testing done over the two hottest days of the summer 2018 – 26^{th} and 27^{th} of July 2018 (see Supplements S1). We thereby used the WRF Single–moment 3–class microphysics scheme (Hong et al. 2004), the Dudhia shortwave and RRTM longwave schemes (Dudhia 1989; Mlawer et al. 1997), and the revised MM5 surface layer scheme (Jiménez et al. 2012). In the first domain, the Kain–Fritsch convection scheme was activated (Kain 2004) and then turned off in the second and third domains, which were at convection-permitting scales. We set the model top at 50 hPa with an additional 5000 m damping layer and subdivided the atmosphere into 56 vertical layers. We used the Noah-MP land surface scheme (Niu et al. 2011; Yang et al. 2011) in its default parameterization over 4 soil layers.

Urban canopy parameters required by the WRF BEP-BEM model were provided via the newly standardized WUDAPT-TO-WRF (W2W) python package developed by Demuzere et al. (2021), following the Fortran version used by Brousse et al. (2016). This allowed the transfer of spatially-explicit morphological urban canopy parameters suitable for urban climate simulations via Local Climate Zones (LCZ) maps covering the inner domain (Figure 2, lower panel). We use the European LCZ map by Demuzere et al. (2019). Thermal and radiative parameters are also directly derived from the LCZ classification and follow those used by Stewart et al. (2014), who used these parameters for the city of Basel, Switzerland. Each parameter for roofs, walls and roads is related to each modal LCZ of the 1 km grid cell via the URBPARM_LCZ.TBL (see Table 1). We decided to keep the roughness length for momentum and the lower boundary for temperatures of roofs, walls, and roads identical across each LCZ. We fixed the roughness length at 1.00E-4 m for walls and at 0.01 m for roofs and roads, respectively. For the boundary temperatures, we set it at 299 K for the roofs and the walls, respectively, and at 293 K for the road. We chose to deactivate the air conditioning in our simulation because air conditioning systems are not common in residential areas across London and surrounding cities, which compose the major part of the land use land cover.

In this study, two potential planetary boundary layers (PBL) schemes are compared in terms of performance and need of bias correction: the commonly used Bougeault-Lacarrère scheme (BouLac; Bougeault and Lacarrere (1989)) for urban simulations that use BEP-BEM, and the recently coupled YSU scheme to BEP-BEM (Hong et al. 2006; Hong and Kim 2008; Hendricks

TABLE 1. Thermal and radiative parameters per LCZ based on Stewart et al. (2014). Road parameters are considering a mixture of asphalted and concrete road pavements and grass.

	Heat capacity $[J \cdot m^{-3} \cdot K^{-1}]$			**Thermal conductivity** $[J \cdot m^{-1} \cdot s^{-1} \cdot K^{-1}]$			**Albedo**			**Emissivity**		
	Roof	**Wall**	**Road**	**Roof**	**Wall**	**Road**	**Roof**	**Wall**	**Road**	**Roof**	**Wall**	**Road**
LCZ 1	1.80E+06	1.80E+06	1.75E+06	1.25	1.09	0.77	0.13	0.25	0.15	0.91	0.90	0.95
LCZ 2	1.80E+06	2.67E+06	1.65E+06	1.25	1.50	0.73	0.18	0.20	0.16	0.91	0.90	0.95
LCZ 3	1.44E+06	2.05E+06	1.63E+06	1.00	1.25	0.69	0.15	0.20	0.18	0.91	0.90	0.95
LCZ 4	1.80E+06	2.00E+06	1.54E+06	1.25	1.45	0.60	0.13	0.20	0.20	0.91	0.90	0.95
LCZ 5	1.80E+06	2.00E+06	1.50E+06	1.25	1.45	0.62	0.13	0.25	0.20	0.91	0.90	0.95
LCZ 6	1.44E+06	2.05E+06	1.47E+06	1.00	1.25	0.60	0.13	0.25	0.21	0.91	0.90	0.95
LCZ 7	2.00E+06	7.20E+05	1.38E+06	2.00	0.50	0.51	0.15	0.20	0.24	0.28	0.90	0.92
LCZ 8	1.80E+06	1.80E+06	1.80E+06	1.25	1.25	0.80	0.18	0.25	0.17	0.91	0.90	0.95
LCZ 9	1.44E+06	2.56E+06	1.37E+06	1.00	1.00	0.55	0.13	0.25	0.23	0.91	0.90	0.95
LCZ 10	2.00E+06	1.69E+06	1.49E+06	2.00	1.33	0.61	0.10	0.20	0.21	0.91	0.90	0.95


Domain nesting in WRF (12 km - 3 km - 1 km)

Official and personal weather stations over urban LCZ in Domain 3

FIG. 2. Domain nesting (upper) and urban land cover in the inner domain (lower). The WRF nesting strategy consists of three nested domains at 12 km (D1), 3 km (D2) and 1 km (D3) horizontal resolution. The altitude is plotted to highlight the flat terrain of south-east England covered in D3. In the lower panel, the resulting urban landcover in D3 after using the WUDAPT-TO-WRF python tool is presented in the form of Local Climate Zones (LCZ). The MIDAS official automatic weather stations (AWS) and the Netatmo personal weather stations (PWS) used for the evaluation of the model and the subsequent bias-correction using PWS only are overlayed in grey. The sea is shown in blue in the lower panel while coastlines are drawn in black in the upper panel.

et al. 2020). Although we found that the latter performed better over the two hottest days of summer 2018 (see Appendix A), we decided to keep a simulation with BouLac as YSU has only been applied over Dallas (Wang and Hu 2021) whereas BouLac has been used in multiple studies already (e.g., Salamanca et al. (2011), Salamanca et al. (2012), Gutiérrez et al. (2015), Tewari et al. (2017), Mughal et al. (2019)). The Mellor-Yamada-Janjic (MYJ; Janjić (1994), Janić (2001)) scheme, also available for BEP-BEM simulations, is disregarded in this study since this PBL scheme is especially used for mountainous terrain (Zonato et al. 2022), and we are modelling the relatively flat terrain of south-east England.

b. Model evaluation

We evaluate the model's performances against 35 official weather stations' measurements of air temperature at 2 m obtained from the UK Met Office MIDAS network (Sunter (2021), UKMO (2021); Figure 1, lower panel). To address the issue of lack of official observations amongst the urban environment, we use Netatmo PWS to complement the model evaluation (Figure 1, lower panel). Prior to the evaluation, unrealistic PWS measurements were filtered out using the Crowd-QC v1.0 R package from Grassmann et al. (2018); details of the method can be found in Napoly et al. (2018) and other publications such as Brousse et al. (2022), who used the same dataset over London. This resulted in 407 urban PWS suitable for evaluating the UCM. Each model simulation is evaluated using a set of common statistical indicators: the root mean squared error (RMSE), the mean absolute error (MAE), the mean bias error (MB), Spearman's coefficient of correlation (r) and the square of Pearson's coefficient of correlation (r^2). These metrics are obtained using the Python scikit-learn and scipy's stats packages from Pedregosa et al. (2011) and Virtanen et al. (2020).

c. Bias correction using personal Netatmo weather stations

We expect urban climate simulations to have systematic biases that can be induced for a variety of reasons, such as: urban canopy parameters (Demuzere et al. 2017; Hammerberg et al. 2018; Zonato et al. 2020); complexity of urban climate models (Grimmond et al. 2011; Loridan and Grimmond 2012; Lipson et al. 2021); time at which the simulation is initialised (Bassett et al. 2020); choice of initial and boundary conditions for lateral and vertical forcing (Brisson et al. 2015); or choice of model parameterizations – such as the two evaluated in this work. Hence, UCM will always present a certain degree of uncertainty that has to be coped with prior to performing urban climate impact studies that use climatic variables derived from modelled simulations to estimate the impact of the urban climate on other events (e.g. mortality, biodiversity, etc.). In our study, we decided to use machine learning regressors and benefit from the high density of PWS in south-east England to correct the air temperature biases and make the simulations usable for urban heat impact studies. To our knowledge, such technique has never proposed as a viable approach for bias-correction of urban climate simulations, probably because of the lack of observations in urban areas.

Indeed, the practice of bias correction is to find a transformation between modelled variables and measured variables. Common practices include adding the mean bias to the modelled variable distribution or applying a separate correction to each quantile of the distribution (Maraun and Widmann 2018). Typically, observations are available only from official weather stations, which may not capture spatial variation within an urban area. Here, because we want to use observations which represent the spatial variation within urban areas at the 1 km scale, we developed an innovative method for bias-correction. Using regression, we predict the bias in the modelled air temperature at 2 m (T2) relative to the PWS observations at each model grid cell which has PWS observations. This prediction is based on the same set of spatially explicit morphological urban canopy parameters at 1 km horizontal resolution that were inputs to the UCM. These include the urban fraction, the surface height, the average building height, the building surface to plan area fraction (λb), the plan area fraction (λp) and the frontal area fraction (λf). We are therefore making the assumption that the spatial variation in the bias of the model is dependent only upon its spatial inputs.

We chose to bias-correct the simulated daily minimum, maximum and average T2 ($T2_{min}$, $T2_{max}$, and $T2_{mean}$) using filtered PWS observations in London and south-east England. To do so, only PWS that have less than 4 hours per day without data and that are located in urban pixels with an urban fraction greater than 0 are retained – where the WRF land-use land-cover at 1 km horizontal resolution refers to an LCZ. Daily temporal scale is considered optimal as it combines a higher spatial density of measurements compared to hourly data and a lower computational requirement; it is also a commonly used temporal scale for urban heat impact studies. Daily minimum and maximum air temperatures at 2 m are defined following the Met Office Had-UK definition: minimum temperature observed from 9AM of the previous day d-1 to 9AM of the d

Table 2. Hyperparameter tuning used by each regressors

Model	**Parameters Dictionary**
Linear	’normalize’: False
Ridge	’alpha’: 1, ’normalize’: True, ’random_state’: 42, ’solver’: ’lsqr’, ’tol’: 0.01
Lasso	’alpha’: 1, ’normalize’: False, ’random_state’: 42, ’selection’: ’random’, ’tol’: 1e-10
Random Forest	’max_features’: ’sqrt’, ’min_samples_leaf’: 11, ’min_samples_split’: 2, ’n_estimators’: 400, ’random_state’: 42
Gradient Boosting	’learning_rate’: 0.2, ’max_depth’: 3, ’max_features’: ’sqrt’, ’min_samples_leaf’: 10, ’min_samples_split’: 22, ’n_estimators’: 200, ’random_state’: 42, ’subsample’: 0.2

day, and maximum temperature observed from 9AM of the d day to 9AM of the next day d+1 (Hollis et al. 2019).

We test the ability of 6 different regressors of increasing complexity available in the Python scikit-learn packages (Pedregosa et al. 2011) to predict the model bias based on WRF spatial urban canopy parameters only. These regressors are: dummy regression (which simply returns the mean), linear regression, Ridge regression, Lasso regression, Random Forest regression, and Gradient Boosting regression. Each of the different regressors, except the dummy regression, offers a set of parameters that can be fine-tuned to increase each regressor’s performance. Hence, prior to running the daily bias-correction we use a 5 K-fold cross-validation using the Grid Search CV package from scikit-learn in Python to evaluate the impact of hyperparameter tuning on the regressors’ performances based on RMSE, MAE and r^2. The cross-validation is done over the summertime average daily mean temperature bias from the control run only, for computational reasons. We retain RMSE as the refitting score to better capture the spatial spread and extremes of T2. The resulting parameterizations are given in Table2. We chose to keep the same hyperparameter tuning for all bias correction and predictions to ease comparability between the outcomes.

Once the hyperparameter tuning is done and prior to performing the final bias-correction, we test if the bias-correction is beneficial for palliating to the models’ bias and if it also benefits from training the regressors at the daily time-step or if a training using the time-mean bias is sufficient. To perform this evaluation using the same metrics as in the model evaluation, we bootstrap each regressors 25 times, randomly sampling 80 % of the data as training and the remaining 20 % as testing – for both the daily-minimum, -maximum and -average, and their respective summer time-mean average. The average predicted daily T2_BC of all predicted T2_BC in the test sample is then compared against the observed T2 – for daily-minimum, -maximum and -average.

After this final step, we bias-correct both the BouLac and the YSU runs using 100 % of the PWS data to compare the spatial outcomes. We also predict T2 out of PWS' observed T2 with the same set of covariates used to predict the model bias to illustrate how divergent each bias-corrected model outputs are to a simplified predicted T2 that is not a derivative of any model constraint. Because more refined and complex techniques exist to predict air temperature from PWS and very high-resolution earth observations (e.g., Venter et al. (2020), Venter et al. (2021)), we do not evaluate these predicted temperatures which should simply be considered as an illustration of how bias-corrected products are similar or divergent to observational data.

Lastly, to illustrate the potential benefit of modelled air temperature bias-correction prior to urban heat impact studies, we calculate the average population weighted temperatures – based on the United Kingdom census data from 2011 – in Greater London before and after the bias-correction.

3. Results

a. WRF simulation evaluation

When we evaluate the two model simulations against MIDAS official weather stations only, they perform similarly, demonstrating a systematic negative bias of ~0.55 °C on average (Table 3). The average correlation with the automatic weather stations following the squared Pearson's r^2 is of 0.77 for BouLac and 0.79 for YSU, while using Spearman's r it is of 0.86 and 0.88, respectively. A slight decreased performance is found in urban pixels for YSU, with an average MAE of 1.83 °C and a negative MB of 0.79 °C compared to BouLac's 1.82 °C for MAR and -0.56 °C for MB. In general, the bias is more important at night, and, in non-urban stations, performances are similar. Hence, looking only at the models' performances using standard in-situ observations doesn't provide information on which model represents the urban climate more accurately.

On the other hand, comparison with PWS observations identifies differences in performance in urban areas between the models, as shown by the performance metrics plotted in Figure 3 and B1. The BouLac simulation has a stronger cool bias of -1.46 °C ± 0.6 °C on average in the urban area, compared to YSU's MB of -0.99 °C ± 0.82 °C. RMSE and MAE are similar, with values of 2.79 °C ± 0.36 °C and 2.20 °C ± 0.32 °C for BouLac and 2.66 °C ± 0.40 °C and 2.15 °C ± 0.35 °C for YSU. These metrics are consistent with the MIDAS observations, highlighting a systematic cool bias of the model and a coefficient of determination (r^2) of 80 %. Importantly, the variability in

TABLE 3. Average of all performance metrics calculated at each MIDAS official weather stations for hourly air temperature at 2 m for the summer period (1^{st} June 2018 to the 31^{st} of August 2018). Urban stations are stations located in a pixel classified as an urban LCZ in WRF and rural stations are located in other natural land-use land-cover.

	BouLac					**YSU**				
	RMSE	**MAE**	**MB**	**r^2**	**r**	**RMSE**	**MAE**	**MB**	**r^2**	**r**
All	2.33	1.82	-0.56	0.77	0.86	2.31	1.83	-0.57	0.79	0.88
Urban	2.42	1.88	-0.73	0.76	0.86	2.42	1.92	-0.93	0.77	0.87
Rural	2.32	1.81	-0.53	0.78	0.86	2.28	1.81	-0.50	0.80	0.88

the model's performance is more greater in the YSU run – reflected by greater standard deviations of performance metrics – and, in the BouLac simulation, the metrics are more heterogeneously distributed amongst the urban area. Indeed, when we look at the YSU simulation, we can see that the model has a smaller MB in suburban areas and a greater MB in the city centre. Yet, in parallel, the correlation with the PWS is lower in the suburban areas and higher in the centre of the city. This could mean that YSU accurately represents the urban temperatures on average due to compensating effects, which we do not intend to evaluate in this study. Nevertheless, this shows how PWS are beneficial for capturing the spatial heterogeneity of each model's performance and therefore supports the use of spatially-varying bias-correction.

b. Bias correction of urban climate simulations

Over our domain of study covering south-east England during the Summer 2018, both models are subject to a cold negative bias of ∼-0.5 °C on average according to official stations and of ∼-1.0 °C to ∼-1.5 °C according to PWS. But as demonstrated above, the bias of the models against PWS observations has substantial spatial variation and so the bias correction for urban heat impact studies should be spatially explicit.

After performing a bootstrapping procedure – running the bias-correction 25 times with 80 % of the CWS as training data and 20 % as testing samples – we can see that each machine learning regressors give similar performance(Figure 4; values numerically given in Tables B1 and B2). All bias-corrections were however beneficial compared to the original outputs from the WRF model, reducing RMSE, MAE and MB by 0.29 °C, 0.32 °C and 1.02 °C on average. The bias-correction was most efficient for daily-minimum temperatures and less for daily-maximum temperatures,

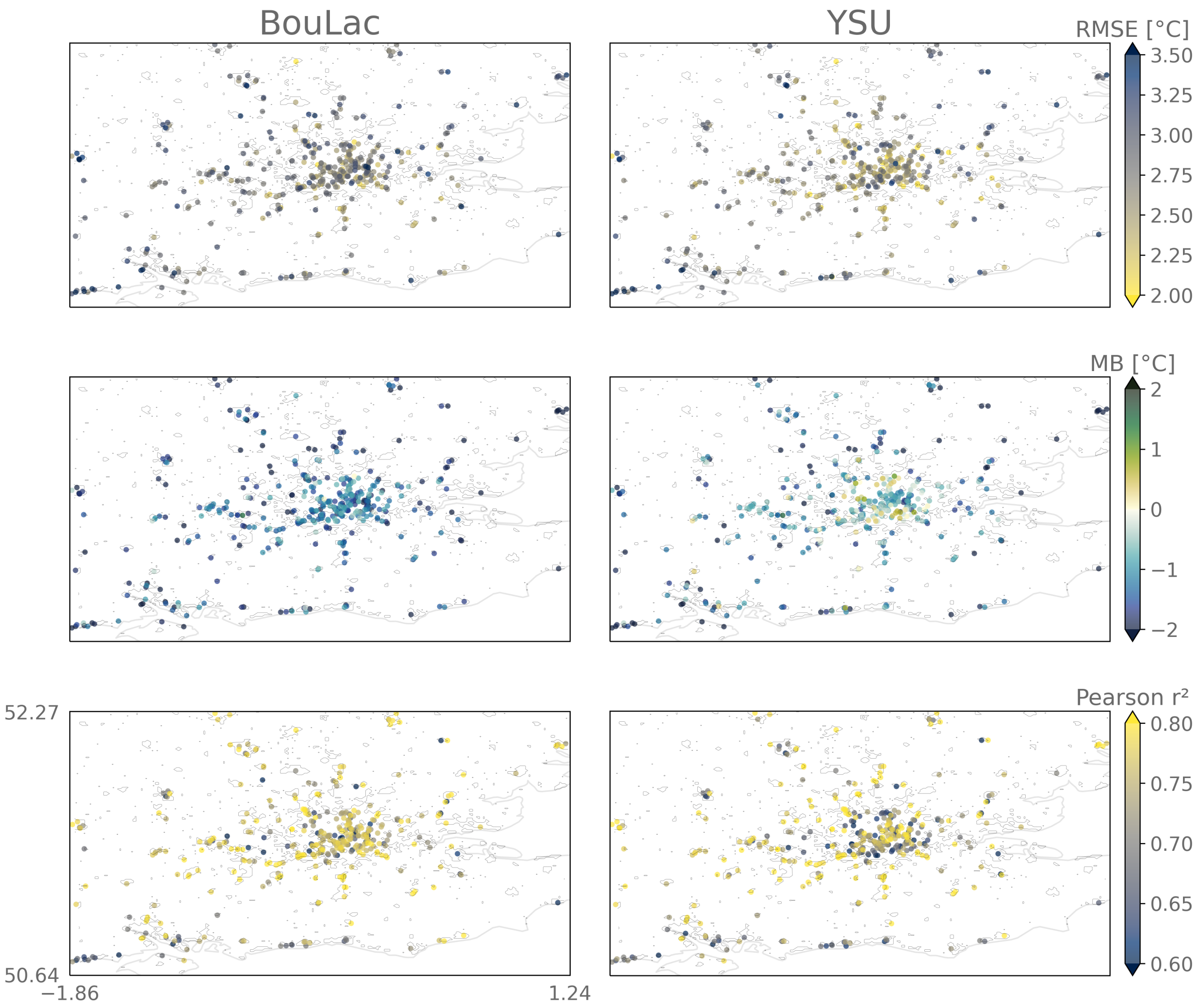


FIG. 3. Performance metrics calculated at location of each citizen personal weather station (PWS) for the two model simulations using different planetary boundary layer schemes (YSU and BouLac). The metrics are calculated over the whole summer 2018 with hourly outputs of near surface air temperature at 2 m. Root mean square error (RMSE) and mean bias (MB) are given in degrees Celsius (°C). Coeffecients of correlation measured with the squared Pearson's r are also provided. Mean absolute error (MAE) and Spearman's r are given in Figure B1 to increase clarity.

where RMSE was not diminished – if not slightly increased (by 0.05 °C for YSU daily-maximum temperatures for example) – by the time-step bias-correction. Interestingly, the spatial correlation between the bias-corrected and the observed temperatures are low, with values ranging from around

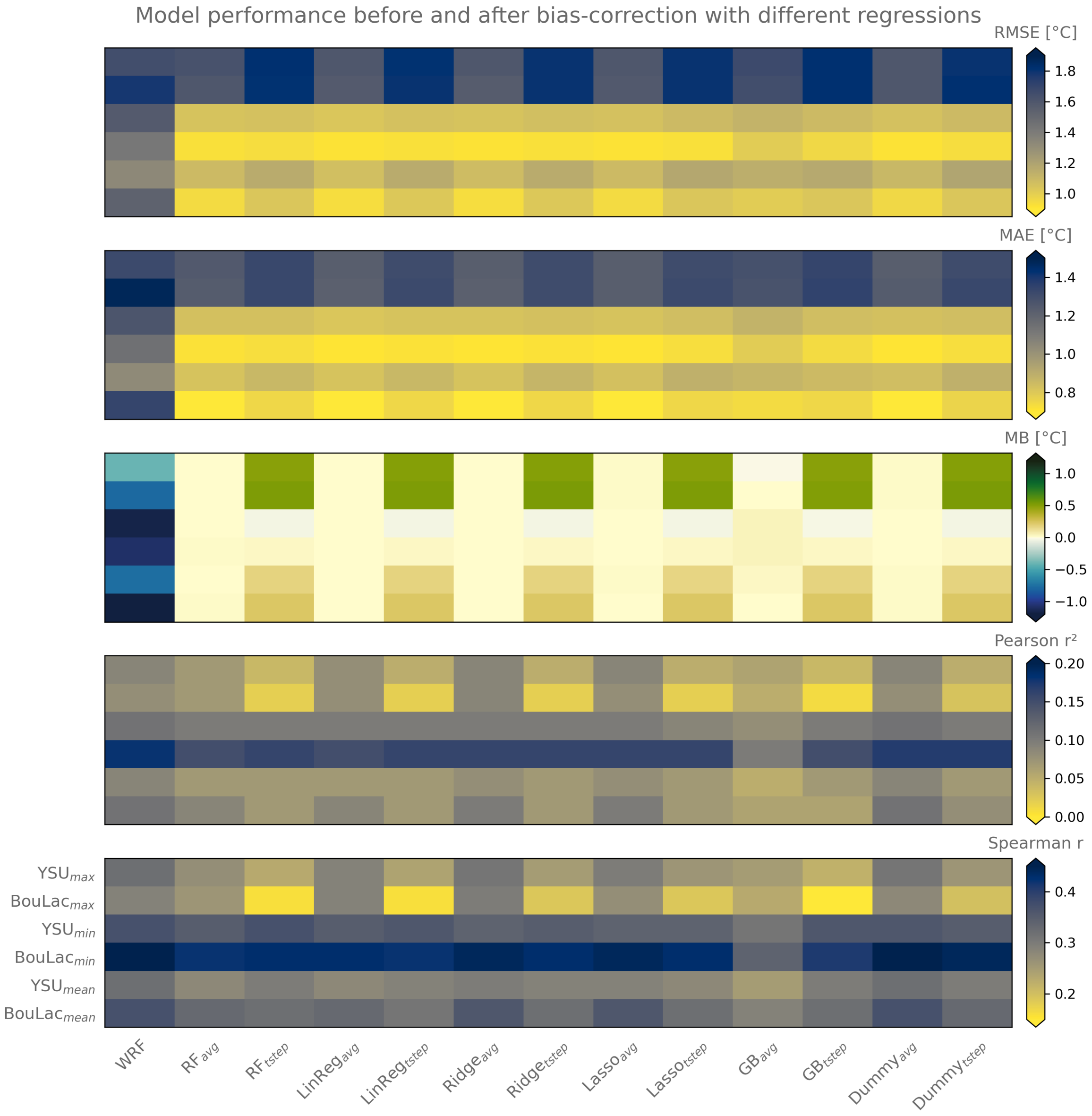


Fig. 4. Performance metrics for the model prior to the bias-correction (WRF) and all the different regressions (random forest: RF; linear regression: LinReg; Ridge regression: Ridge; Lasso regression: Lasso; gradient boosting: GB; and dummy regression: Dummy). The different regressions are assigned a suffix: "avg" for regressions that were trained on the summer time-mean average of daily-minimum, -mean or -maximum temperatures, and "tstep" for those that were trained with the temperatures at each daily time-step.

0.02 to 0.2 for the squared Pearson's r and from around 0.15 to 0.45 for Spearman's r. This can be expected as machine learning algorithms have difficulties representing a time-varying variable with static spatial elements only (Georganos et al. 2021; Venter et al. 2021). Unexpectedly, we find that the training at the daily time-step does not outperform the training at the summer time-mean in terms of spatial correlation with the heat distribution across London. Nonetheless, if we take the average daily-minimum, -mean and -maximum temperatures of all CWS and compare it to the modelled temperatures, we find that the time-step bias-correction is closer to the observations (Figures B2 to B4).

Comparing the spatial differences of the bias-corrected products related to the complexities of each regressors, we find that although each regressor is performing similarly on average, important disparities are found between the outputs. For example, when looking at the average bias-correction imposed to daily-minimum temperatures after training the regressors at each time-step, the Lasso and the Ridge regressors impose a flat bias-correction, similar to the dummy regression, while the random forest and gradient boosting regressors' degrees of freedom result in a spatially diverse bias-correction (Figure 5 and Figures B5 and B6). Besides, the linear regression imposes an average bias-correction spatially-correlated to the modal LCZ. In general, the signal is consistent across each regressors, apart from the Lasso and the dummy regression, where, for YSU, central London requires a stronger bias-correction by 1 °C to 2 °C °C compared to the suburban areas where the bias-correction is around 0.5 °C ; for BouLac, the central bias-correction is lower than YSU. We find that these spatial tendencies are also found for daily-maximum and daily-average temperatures, defending our hypothesis of a systematic bias correlated to spatially explicit input parameters. The spatial differences in bias-correction are however less important for daily-maximum temperatures, which is the time at which the urban heat island is also expected to be the lowest.

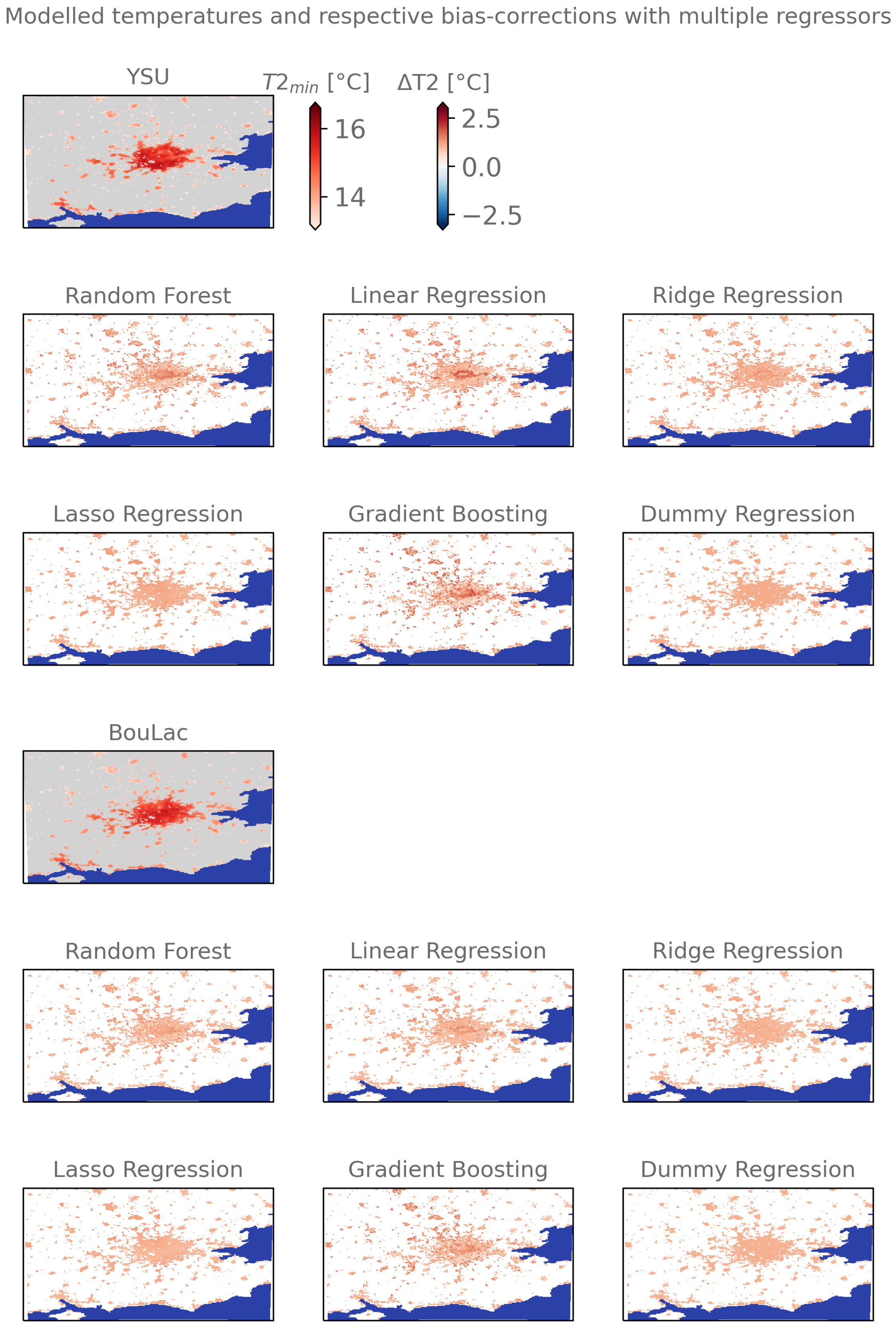

Modelled temperatures and respective bias-corrections with multiple regressors
YSU
$T2_{min}$ [°C]
16
14
ΔT2 [°C]
2.5
0.0
−2.5
Random Forest
Linear Regression
Ridge Regression
Lasso Regression
Gradient Boosting
Dummy Regression
BouLac
Random Forest
Linear Regression
Ridge Regression
Lasso Regression
Gradient Boosting
Dummy Regression

FIG. 5. All regressions propose different bias-corrections (ΔT2) of the average modelled absolute daily minimum urban temperature ($T2_{min}$). Differences of bias-correction are observed between the runs with different planetary boundary layer schemes (Bougeault-Lacarrère – BouLac, and Yonsei Universiy – YSU). The centre of London is subject to a stronger bias-correction. Rural lands are masked in grey and the seas in blue. Bias corrections of daily mean and maximum temperatures are given in Figures B5 and B6

Finally, we find that the bias-corrected BouLac simulation corresponds spatially to predicted temperatures using PWS more than YSU – something we find equally across all regressors (Figure 6 and Figures B7 to B11). As an example, when comparing the average bias-corrected products using the time-step trained random forest regressor we can see that YSU urban heat is more homogeneously distributed than BouLac's or the predicted temperatures from PWS only. BouLac's bias-corrected product shows stronger urban heat in central London compared to suburban areas, coherent with the predicted temperatures. Nonetheless, BouLac's suburban areas are hotter by 0.5 °C to 1.0 °C than the predicted ones with PWS only. This remains less pronounced than in YSU. Lastly, we can see that both bias-corrected products show similar trends when compared to the PWS-only predicted temperatures with hotter suburban areas and cooler secondary cities as well as coastlines. Again, this does not show which product between the PWS-only predicted temperatures and the bias-corrected products is better since we do not evaluate this here.

These results show that bias-correction of modelled air temperature change their spatio-temporal distributions. When focusing on the potential impact bias-correction may have in estimated urban heat impact on urban health, we find that using the random forest regression trained at each daily time-step leads to an increased average population weighted temperature by 0.77 °C in the YSU case, and of 1.24 °C in the BouLac case. Raw model outputs are thereby lowering the impact of heat on the urban population.

4. Discussion and conclusions

In this study, we argue that the joint use of crowd-sourced personal weather stations (PWS) and urban climate models (UCMs) can add value to urban climate research and in particular to urban climate impact research. This is supported by two major outcomes of our case-study focused over London during the summer 2018. First, we showed that evaluation of urban climate simulations using PWS enables the detection of spatially-varying systematic biases in urban areas related to the

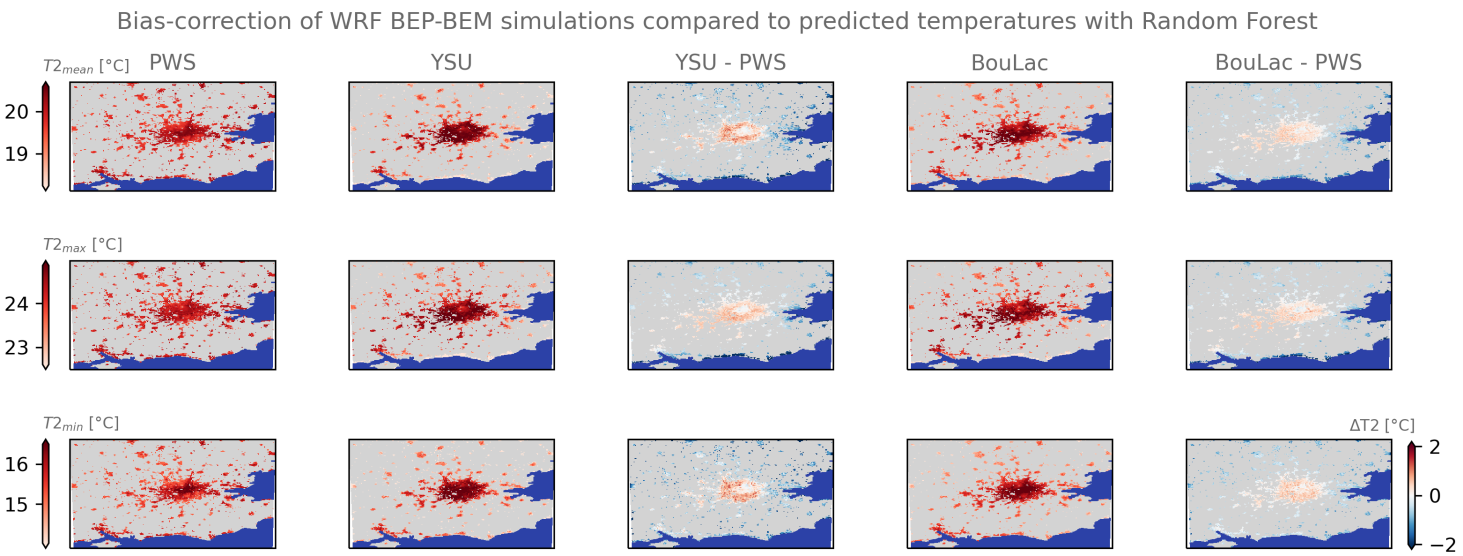


Fig. 6. The random forest regressor leads to different bias-corrections of the two WRF simulations parameterized with different turbulence schemes – the Yonsei University (YSU) and the Bougeault-Lacarrère (BouLac) – and with the BEP-BEM urban canopy model activated. This holds for average daily mean, minimum and maximum temperatures ($T2_{mean}$, $T2_{min}$ and $T2_{max}$) after the daily time-step bias-correction. Compared to the predicted temperatures using the personal weather stations data only (PWS), the bias-corrected products are hotter in the suburban areas of the Greater London and cooler in the rural areas. The difference is more pronounced in YSU (see YSU – PWS). Greyed areas represent natural areas where the bias-correction is not performed and the sea is shown in dark blue. The same figures for the other regressors are given in Figures B7 to B11

UCMs' parameterization, which are not detectable using only official weather stations. Second, we demonstrated that PWS, combined with detailed morphological data derived from LCZ maps, can be used to derive a spatially varying bias-correction via commonly used machine-learning regressors. This latter point has major implications for urban climate impact research – and especially future urban climate impact studies – as we hereby propose the first bias-correction technique that considers the existence of a non-linear spatially heterogeneous bias in modelled urban climates.

Of course, using PWS for evaluating UCM simulations should always cautiously be considered because of the lower accuracy of PWS and the potential uncertainties related to user-driven mistakes in the set-up of their PWS (e.g., indoor sensors instead of outdoor, poor shading conditions, height of the sensor, etc.). However, reliable tools have now been developed since the first use of PWS for model evaluation by Hammerberg et al. (2018) to filter dubious measurements out (e.g., Crowd-QC from Napoly et al. (2018) or Crowd-QC + by Fenner et al. (2021)), thus making PWS observations increasingly reliable. This does not resolve the question of the representativity of measurements, i.e., "how is one PWS measurement representative of the simulated urban pixel?" Yet, the increasing density of PWS in the urban environments begins to alleviate this uncertainty. For example, Venter et al. (2020) found that a density of one PWS per square kilometre is optimal for predicting seasonal air temperature in Oslo. Dense PWS networks hence permit the detection of systematic biases that would otherwise pass undetected. Therefore, to support the development of PWS as a source of urban weather observations for model evaluation, urban climate scientists should identify an optimal density of PWS for UCM evaluation, to define which cities are in need of urban weather observations, and to start instigating common frameworks and standards.

We consider our study innovative and supportive of future advances in the field because it is the first bias-correction technique in urban environments which considers that UCMs' simulated UHI is spatially heterogeneous in its accuracy and that the UHI is not solely linearly correlated to the urban fraction. Aided by the expanding fields of crowd-sourcing weather observations through PWS, machine learning, and potentially deep learning, we infer that our work should serve as the basis of future research that would try, but not restricted to, improving the bias-correction of urban climate models using PWS. For instance, we did not find any machine learning regressor to be more efficient at predicting the model bias. This could be explained by the rather restricted set of covariates we used for training the regressors as well as the coarse horizontal resolution of 1 km at which the covariates were aggregated to be consistent with the model's spatial resolution. Higher spatial resolutions and more specific satellite earth observations could be used to improve regressors' performance, following up on the work by Venter et al. (2021), for example. When modelling the near-surface UHI, which is not a model bias, their regressor achieved similar performances as ours, with an RMSE of 1.05 °C and a Pearson's r^2 of 0.23. Although the common use of model's input parameters and earth observations as covariates could be beneficial, a particular attention should be given to the choice of earth observations since these should not be decorrelated to the model's physics and dynamics as the purpose would remain the bias-correction.

Besides, as our results showed by comparing the performance of regressors trained at the daily time-step and with the summer time-mean average, regressors could gain in performance by adding a temporal component to the covariates. Following up on this idea, the recent work by (Zumwald et al. 2021) tried predicting the near-surface air temperature in Zurich for the 30^{th} of June 2019 out of ~650 Netatmo PWS' measurements during the preceding week. Their set of covariates consisted of spatial earth observations as well as 35 meteorological predictors that were all derived from one official automatic weather stations. The latter predictors helped training the model to recognise how the temperature measured at each PWS location was related to the meteorological variables measured at the automatic weather stations. Their predictions at hourly time-steps achieved reasonable performances with RMSEs around 1.70 °C. Bias-correction of UCM simulations could hence be improved by incorporating temporally explicit meteorological observations from official weather stations. Notwithstanding, this would require extensive investigation on the area down to which each official station is representative for training the regressors. More geographically oriented machine learning regressors, like the geographical random forests (Georganos et al. 2021), could also help integrate these spatial heterogeneities for an improved bias-correction.

In general, we support the use of quality-controlled PWS observations for bias-correction of urban climate simulations. As shown in this case study, model outputs prior to any bias-correction could lead to under- or over-estimation of urban heat impact on public health. We indeed find that for the summer 2018 in London, average population weighted temperatures were higher after bias-correcting the model outputs, suggesting higher urban heat related mortality during this period. This simple example shows that bias-correction of urban climate simulations could have important implications for calculating the exposure of urban citizen to heat or estimating the urban heat-related mortality. Although preferring bias-corrected model outputs to predicted urban air temperatures from earth observations for present-day urban heat impact studies is not covered in this study – and must be further explored – we still argue that bias-correction should be done prior to any urban heat impact studies that imply using climate model outputs. This argument is especially valid for future climate projections at urban scale and we encourage future research to investigate how to transfer present urban bias-correction coefficients to simulated future urban climates. Doing so, bias-corrected simulations could help targeting areas where heat mitigation or adaptation strategies could be more beneficial as their efficiency is dependent on their location and scales of implementation (Yang and Bou-Zeid 2019; Broadbent et al. 2022).

Acknowledgments. We personally thank Stefanos Georganos for his help and his comments on machine learning classifiers and regressors. We also thank Daniel Fenner and Fred Meier for their valuable insights concerning data acquisition, filtering and treatment of crowd-sourced citizen weather stations. Lastly, we are grateful to Matthias Demuzere and other committed members of the WUDAPT project for providing the European LCZ map and the python W2W tools. CH is supported by a NERC fellowship (NE/R01440X/1) and acknowledges funding for the HEROIC project (216035/Z/19/Z) from the Wellcome Trust, which funds OB and CS.

OB designed the study and led the conception of the manuscript with the support of CH and CS. OB was responsible for the WRF modelling, the model evaluation and the bias-correction. CS provided support in the python coding and in the statistical analysis for the bias-correction. OK was responsible for technical support of the installation of WRF on the University College London's "Kathleen" and "Myriad" super-computers. AZ and AM offered guidance in the set-up of the WRF model v4.3 and urban heat modelling expertise with SK. All authors contributed to the writing of the manuscript.

The authors declare no conflicts of interest.

Data availability statement. The simulations done in this research were performed using the WRF model v4.3 (`https://github.com/wrf-model/WRF.git`). The related outputs presented in this research and the scripts used to produce them are available upon reasonable request addressed to the corresponding author.

APPENDIX A

Model sensitivity testing over the two hottest days of Summer 2018

Prior to running the 3-months simulation, we tested the model's sensitivity to a set of parameterization to assess which model is the best performing model for the 3-months simulation. We perform the sensitivity in a progressive way; parameters are kept if beneficial, removed if detrimental. We chose to run the simulations over the two hottest days of the summer 2018 with one additional day as spin-up time – from the 25^{th} to the 27^{th} of July 2018 – to see how the model is capable of accurately representing an extreme condition in terms of air temperature at 2 m – tested against official MIDAS automatic weather stations and personal Netatmo PWS. The model was also tested for relative humidity and wind speed at 10 m at MIDAS locations where records were available. All wind-speed measurements are converted from knots to $\mathrm{m{\cdot}s^{-1}}$.

We start from Heaviside et al. (2015) model's parameterization, who simulated the impact of urbanization on the local climate in the West Midlands in England, but supplement the CORINE land-use land-cover by the Local Climate Zones classification instead since Brousse et al. (2016) compared both products and proved the added value of LCZ over Madrid. We chose the work by Heaviside et al. (2015) as a starting point since it also uses the BEP urban climate model, coupled to the WRF model and is one of the only WRF simulations done over England.

From there, our simulations tested: i) the use of YSU, recently coupled to the BEP-BEM model (Hendricks et al. 2020), instead of Bougeault-Lacarrere; ii) the use of the more complex land surface scheme Noah-MP in its default parameterization instead of the default Noah land surface model; iii) the forcing by ERA5 reanalysis data at 25 km horizontal resolution instead of ERA-Interim; iv) the reduction of soil moisture by 50 % and its increase by 200 %, following suggestions provided by Martilli et al. (2021). We chose not to test the impact of urban canopy parameters in this case to keep our simulations standardized and universally coherent through the LCZ scheme. Their simulation used the same micro-, clouds, convection and radiation physics than ours.

We found that all steps taken from the original parameterization by Heaviside et al. (2015) were beneficial to the model's performance. Through an intermediate simulation where we tested again the BouLac turbulence scheme after step iii, we found that YSU was still performing better.

APPENDIX B

Additional Figures and Tables

This section presents all the figures that are not given in the main text.

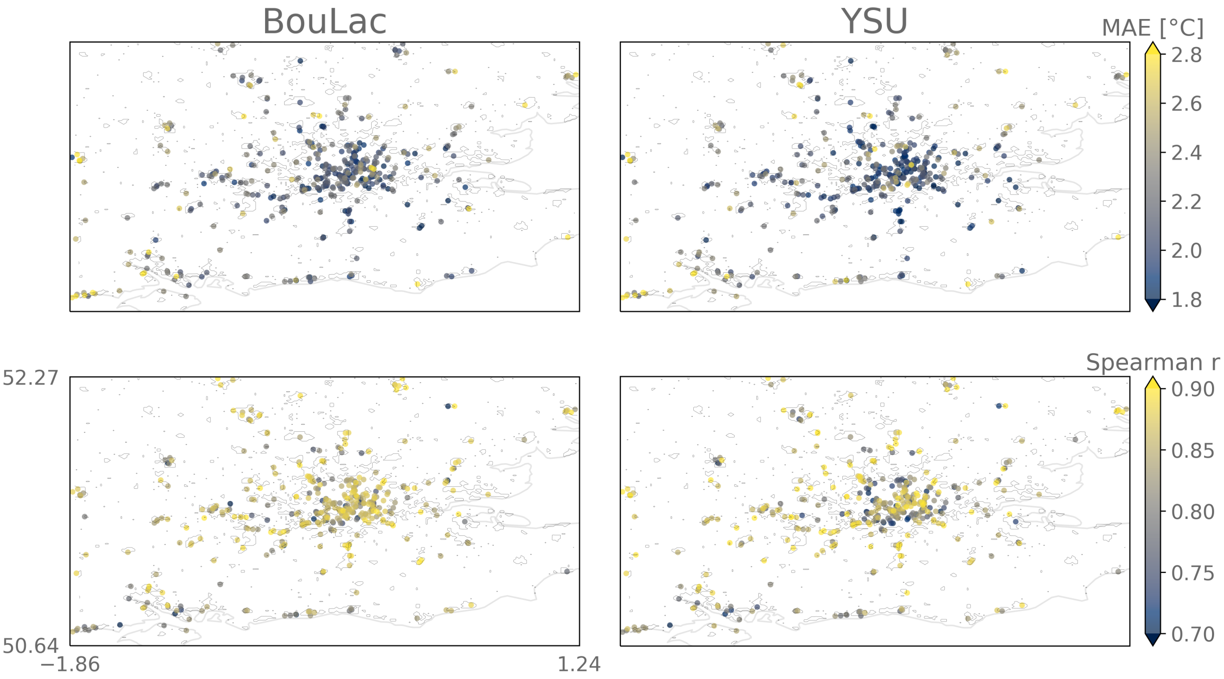


Fig. B1. Same as figure 3, but for MAE and Spearman's r.

TABLE B1. Performance metrics used in Figure 4 for the model using BouLac prior to the bias-correction (WRF) and all the different regressors (random forest: RF; linear regression: LR; Ridge regression: RD; Lasso regression: LA; gradient boosting: GB; and dummy regression: DU). The different regressions are assigned a suffix: “avg” for regressions that were trained on the summer time-mean average of daily-minimum, -mean or -maximum temperatures, and “tstep” for those that were trained with the temperatures at each daily time-step.

BouLac													
	WRF	**RF**$_{avg}$	**RF**$_{tstep}$	**LR**$_{avg}$	**LR**$_{tstep}$	**RD**$_{avg}$	**RD**$_{tstep}$	**LA**$_{avg}$	**LA**$_{tstep}$	**GB**$_{avg}$	**GB**$_{tstep}$	**DU**$_{avg}$	**DU**$_{tstep}$
MEAN													
RMSE	1.54	0.95	1.04	0.94	1.03	0.94	1.03	0.95	1.04	1.01	1.04	0.96	1.04
MAE	1.34	0.69	0.75	0.69	0.75	0.68	0.75	0.69	0.75	0.74	0.75	0.7	0.76
MB	-1.2	0.01	0.23	0	0.23	0	0.23	0	0.23	0	0.23	0.01	0.23
Pearson r^2	0.11	0.09	0.07	0.09	0.07	0.1	0.07	0.1	0.07	0.06	0.06	0.11	0.08
Spearman r	0.37	0.33	0.32	0.33	0.31	0.36	0.32	0.36	0.32	0.29	0.32	0.37	0.33 0.88
MIN													
RMSE	1.42	0.93	0.94	0.92	0.93	0.92	0.93	0.92	0.93	1.01	0.96	0.92	0.94
MAE	1.15	0.72	0.73	0.71	0.72	0.71	0.72	0.71	0.73	0.79	0.74	0.71	0.73
MB	-1.08	0.01	0.02	0	0.02	0	0.02	0	0.02	0.04	0.02	0	0.02
Pearson r^2	0.18	0.15	0.16	0.15	0.16	0.16	0.16	0.16	0.16	0.1	0.15	0.17	0.17
Spearman r	0.46	0.42	0.43	0.43	0.42	0.44	0.43	0.44	0.43	0.34	0.41	0.46	0.44
MAX													
RMSE	1.78	1.6	1.81	1.58	1.8	1.57	1.8	1.59	1.8	1.65	1.82	1.6	1.82
MAE	1.48	1.24	1.33	1.22	1.32	1.22	1.31	1.23	1.32	1.28	1.35	1.24	1.33
MB	-0.79	0	0.52	0	0.52	0	0.53	0.01	0.52	0	0.51	0.01	0.53
Spearman r	0.08	0.07	0.02	0.08	0.02	0.09	0.02	0.08	0.02	0.05	0.01	0.08	0.03
Spearman r	0.29	0.26	0.16	0.29	0.16	0.3	0.19	0.27	0.19	0.23	0.14	0.28	0.2

Table B2. Performance metrics used in Figure 4 for the model using YSU prior to the bias-correction (WRF) and all the different regressors (random forest: RF; linear regression: LR; Ridge regression: RD; Lasso regression: LA; gradient boosting: GB; and dummy regression: DU). The different regressions are assigned a suffix: "avg" for regressions that were trained on the summer time-mean average of daily-minimum, -mean or -maximum temperatures, and "tstep" for those that were trained with the temperatures at each daily time-step.

YSU													
	WRF	**RF**$_{avg}$	**RF**$_{tstep}$	**LR**$_{avg}$	**LR**$_{tstep}$	**RD**$_{avg}$	**RD**$_{tstep}$	**LA**$_{avg}$	**LA**$_{tstep}$	**GB**$_{avg}$	**GB**$_{tstep}$	**DU**$_{avg}$	**DU**$_{tstep}$
MEAN													
RMSE	1.33	1.09	1.16	1.07	1.16	1.08	1.16	1.09	1.18	1.15	1.17	1.1	1.19
MAE	1.04	0.82	0.86	0.82	0.86	0.82	0.87	0.83	0.89	0.87	0.85	0.84	0.89
MB	-0.76	0	0.17	0	0.17	0	0.17	0.01	0.16	0.02	0.17	0.01	0.17
Pearson r^2	0.09	0.07	0.07	0.07	0.07	0.08	0.07	0.08	0.07	0.05	0.07	0.09	0.07
Spearman r	0.32	0.28	0.3	0.28	0.29	0.3	0.29	0.29	0.28	0.25	0.3	0.32	0.3
MIN													
RMSE	1.58	1.05	1.06	1.04	1.06	1.05	1.07	1.06	1.09	1.12	1.09	1.06	1.09
MAE	1.27	0.83	0.83	0.81	0.82	0.82	0.83	0.82	0.84	0.88	0.84	0.83	0.84
MB	-1.17	0	-0.03	0	-0.03	0	-0.03	0	-0.03	0.04	-0.02	0	-0.03
Pearson r^2	0.11	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.09	0.08	0.1	0.11	0.1
Spearman r	0.37	0.35	0.37	0.35	0.36	0.34	0.35	0.34	0.34	0.31	0.36	0.36	0.35
MAX													
RMSE	1.65	1.63	1.82	1.6	1.81	1.6	1.8	1.6	1.8	1.67	1.82	1.6	1.8
MAE	1.32	1.25	1.33	1.23	1.31	1.23	1.31	1.23	1.31	1.29	1.34	1.23	1.31
MB	-0.41	0	0.49	0	0.5	0	0.5	0.01	0.49	-0.01	0.49	0.01	0.5
Pearson r^2	0.09	0.07	0.04	0.08	0.05	0.09	0.05	0.09	0.05	0.06	0.04	0.09	0.05
Spearman r	0.32	0.27	0.23	0.29	0.24	0.31	0.25	0.3	0.26	0.25	0.22	0.31	0.26

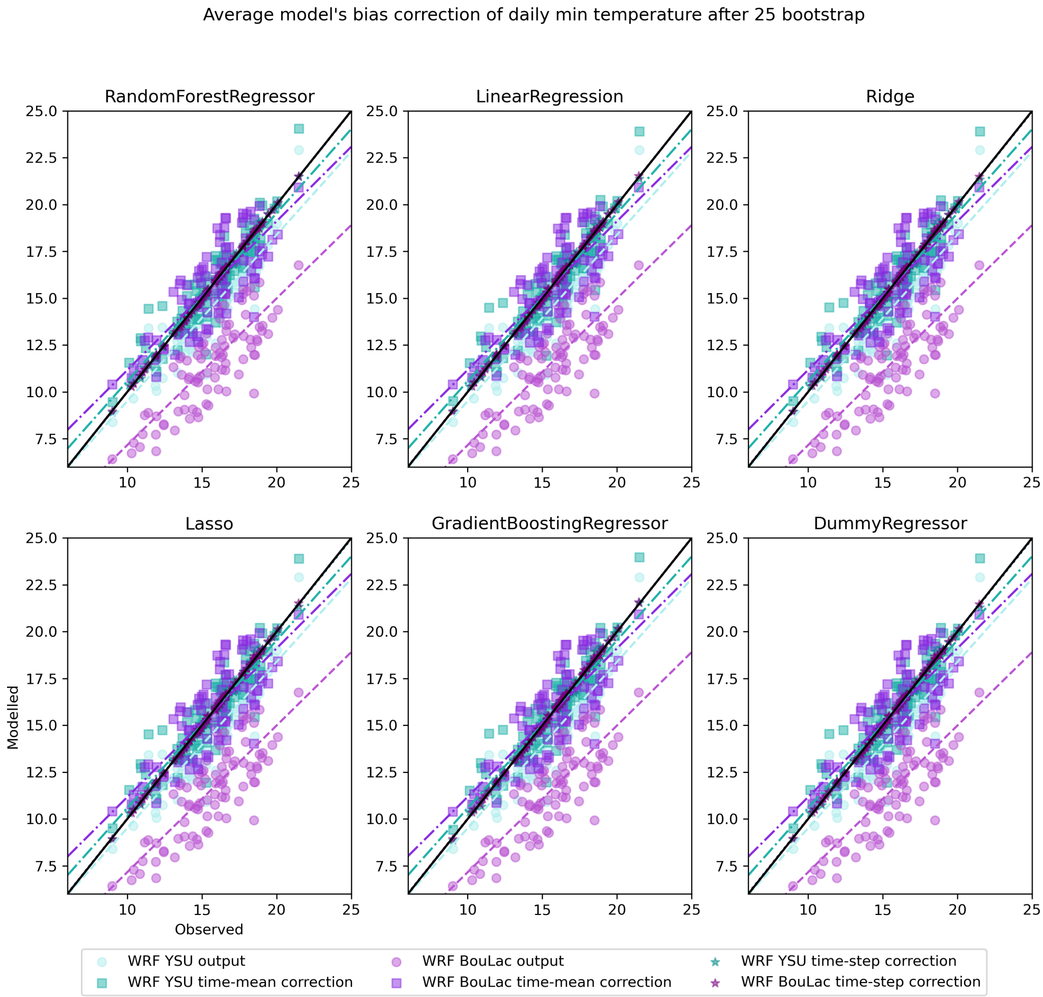


FIG. B2. Average modelled daily minimum air temperature at 2 m against observed at citizens' personal weather stations locations show that all machine learning regressors perform a similar bias-correction on average. In blue, modelled temperatures at 2 m are from the model simulation that used the Yonsei University (YSU) planetary boundary layer scheme before the bias correction (circles), after the summer time-mean bias correction (squares) and after the daily time-step bias correction (stars). In purple, the same values are given for the simulation which used the Bougeault-Lacarrère (BouLac) scheme. Dashed lines represent the least squares polynomial fitted lines and the black full line represents the identity line.

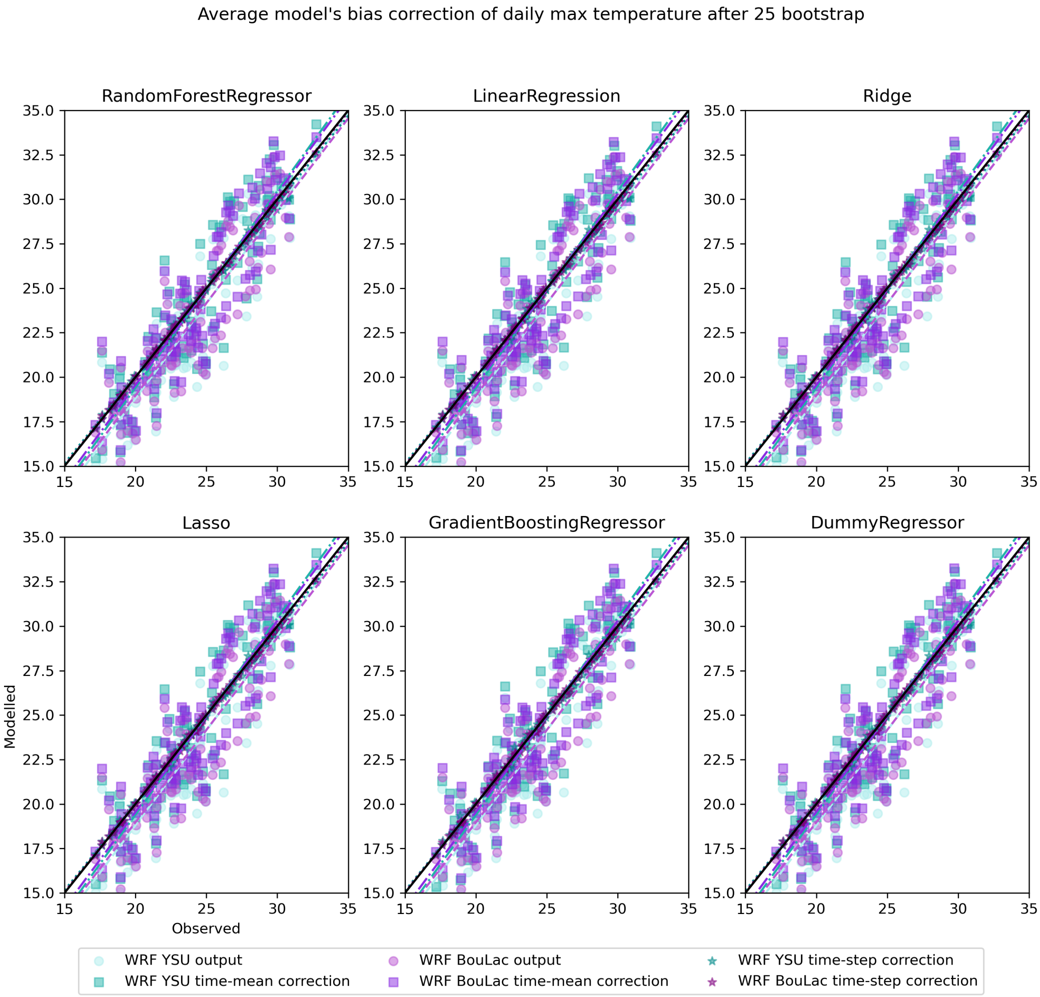


FIG. B3. Same as figure B2, but for daily maximum temperatures.

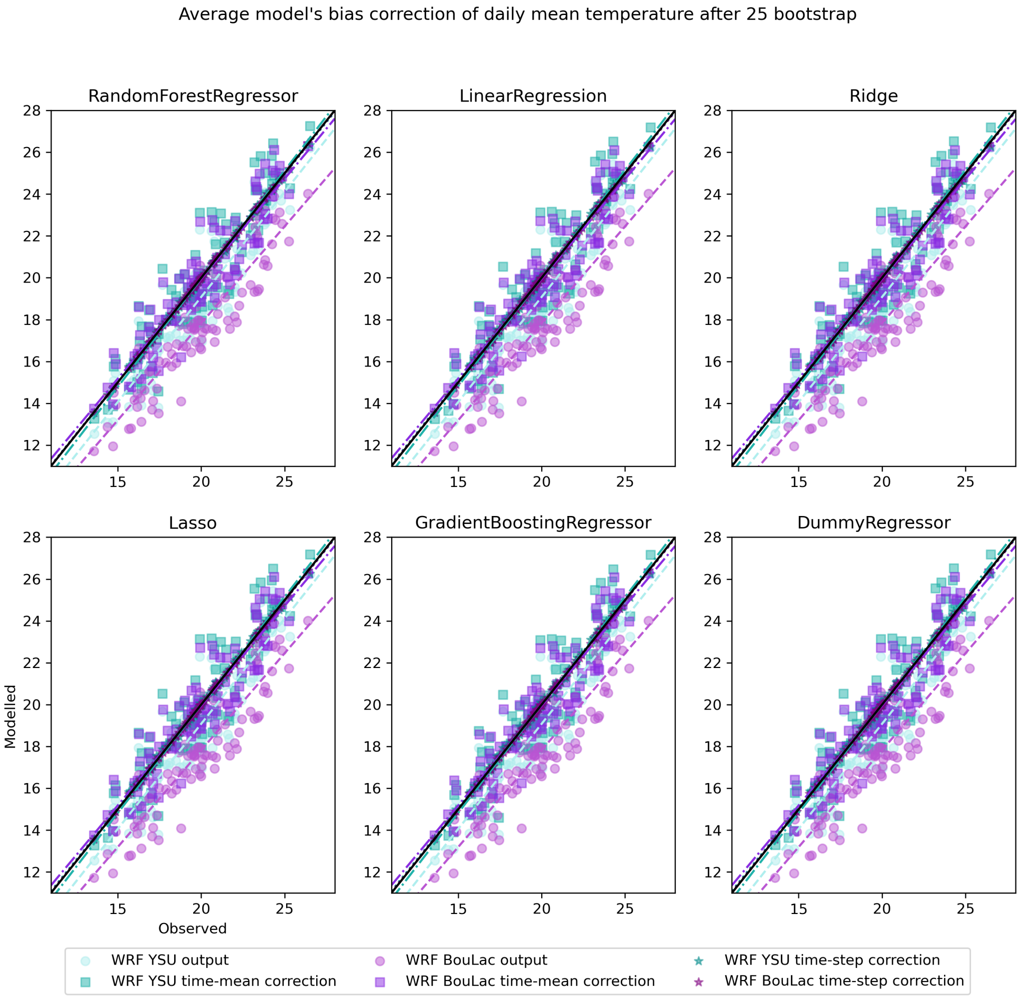


FIG. B4. Same as figure B2, but for daily mean temperatures.

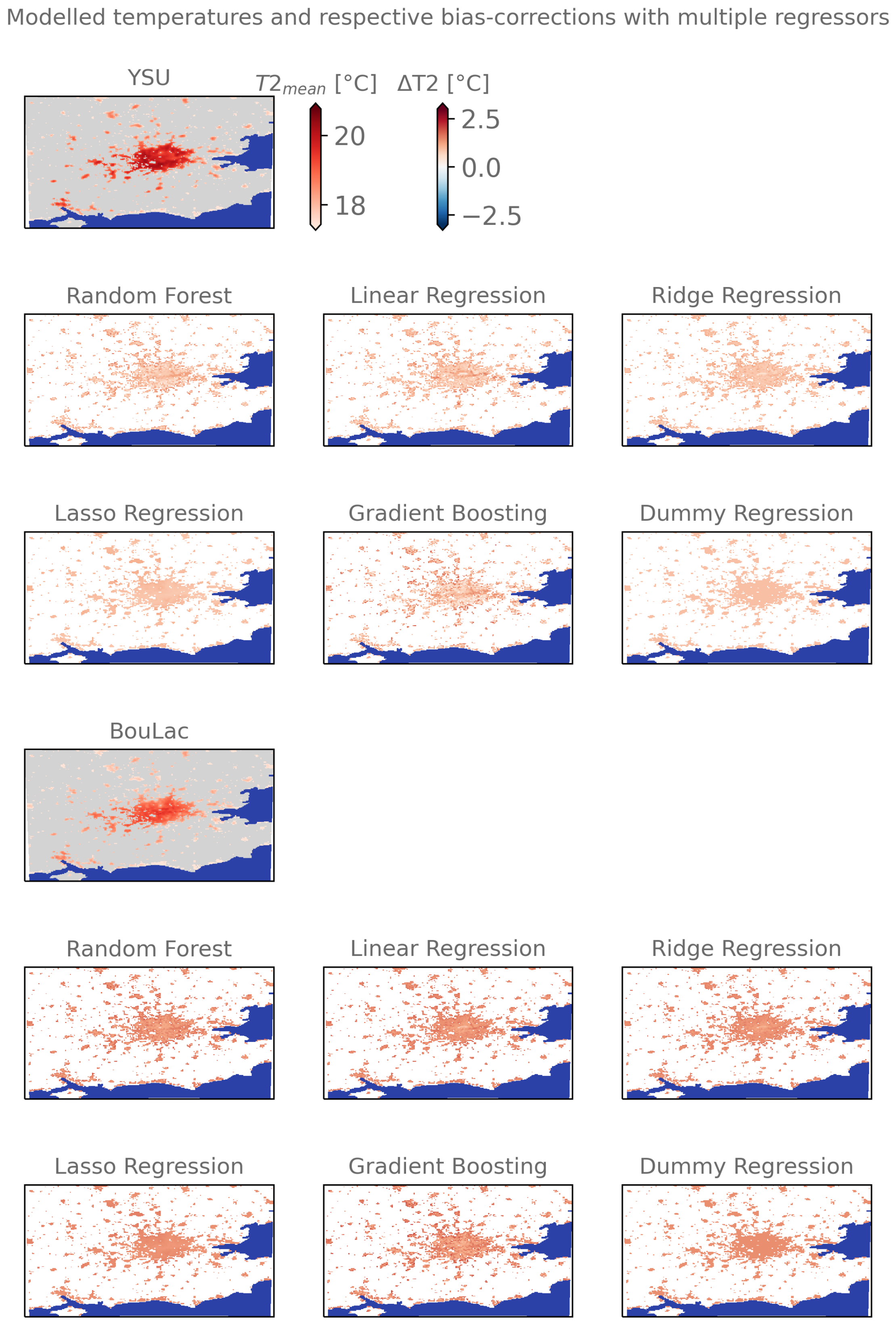


Fig. B5. Same as figure 5, but for daily mean temperatures.

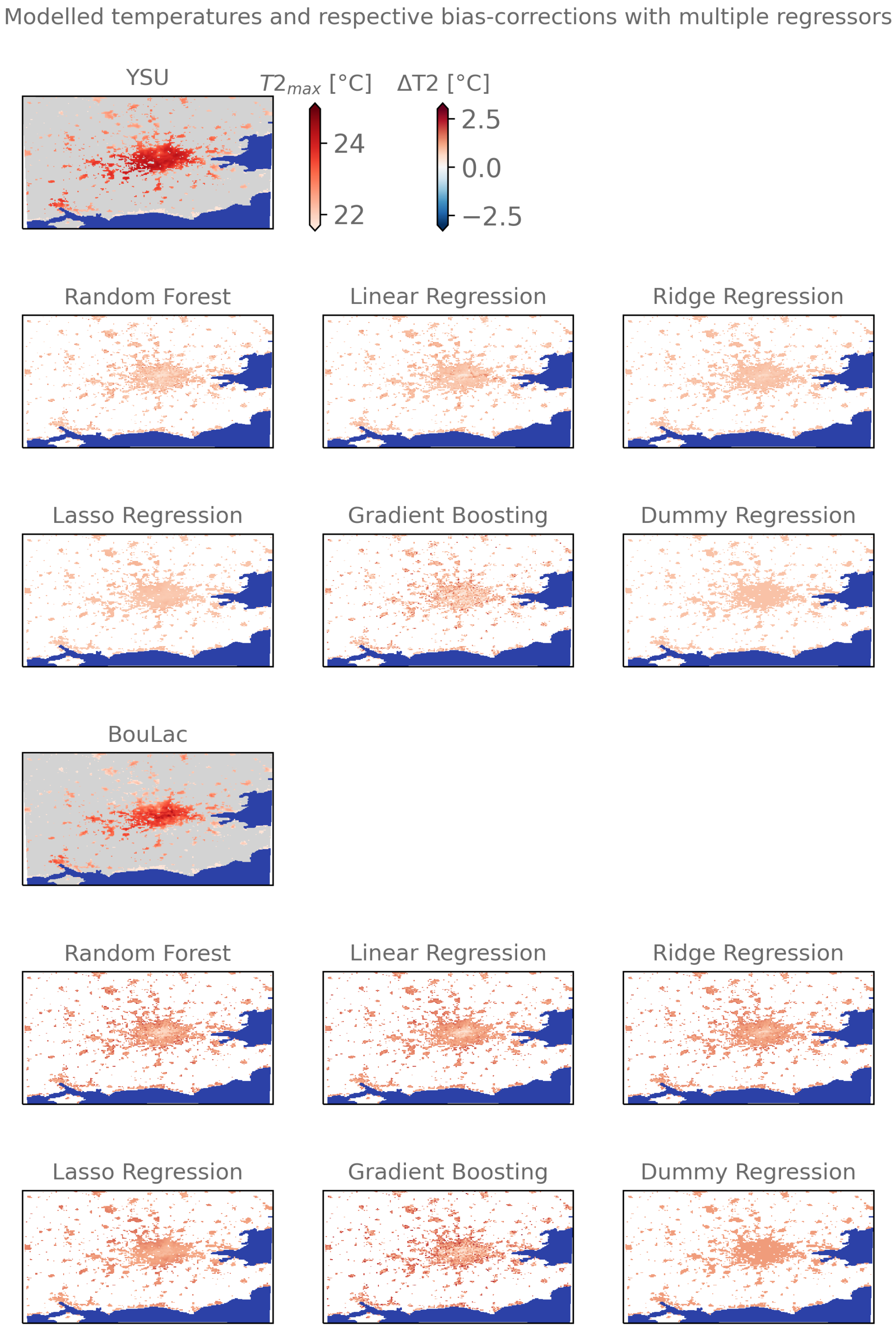


Fig. B6. Same as figure 5, but for daily maximum temperatures.

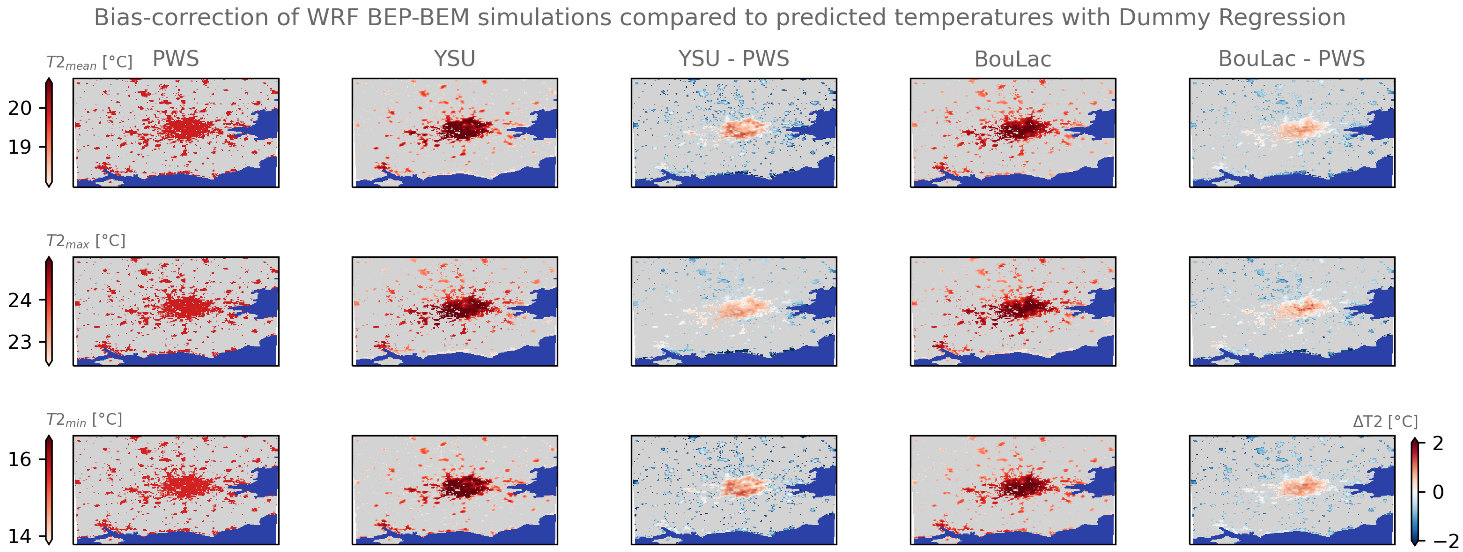


Fig. B7. Same as figure 6, but for dummy regression.

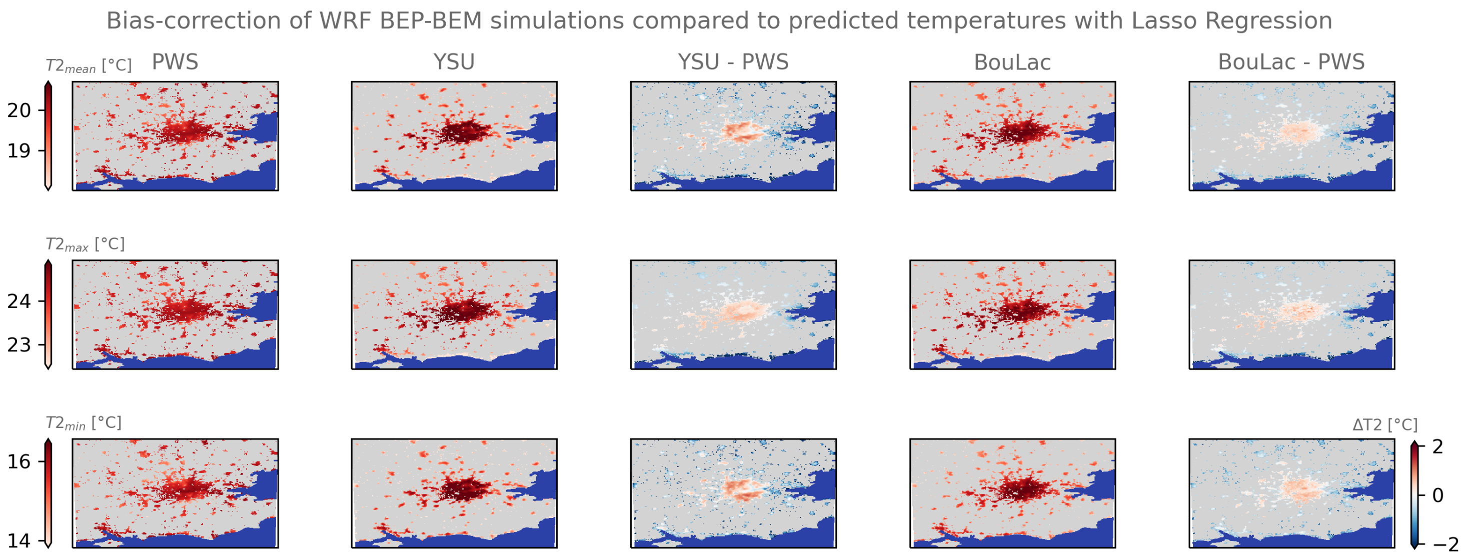


FIG. B8. Same as figure 6, but for Lasso regression.

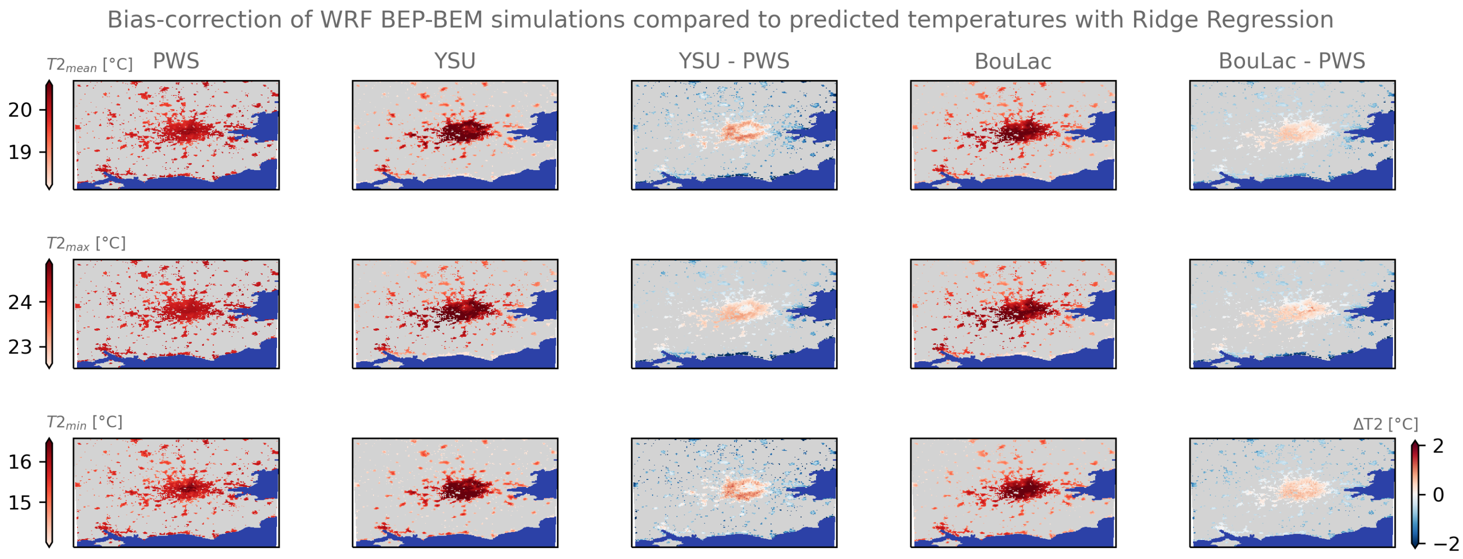


Fig. B9. Same as figure 6, but for Ridge regression.

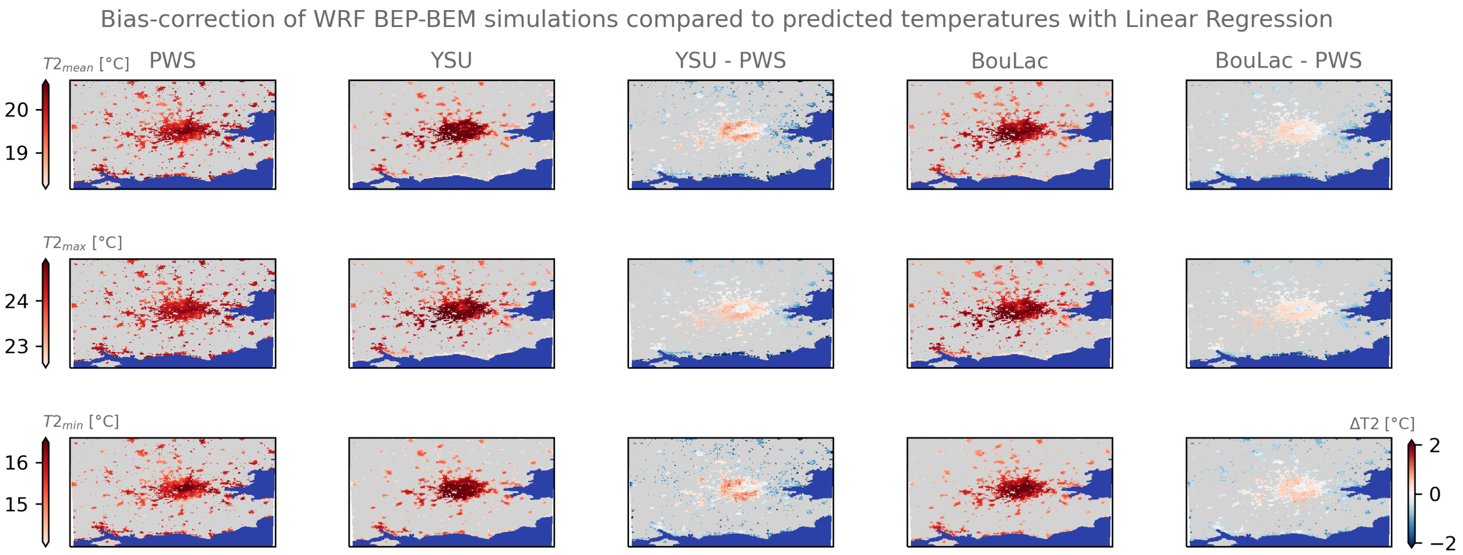

Bias-correction of WRF BEP-BEM simulations compared to predicted temperatures with Linear Regression
PWS
YSU
YSU - PWS
BouLac
BouLac - PWS
T2mean [°C]
20
19
T2max [°C]
24
23
T2min [°C]
16
15
ΔT2 [°C]
2
0
−2

FIG. B10. Same as figure 6, but for linear regression.

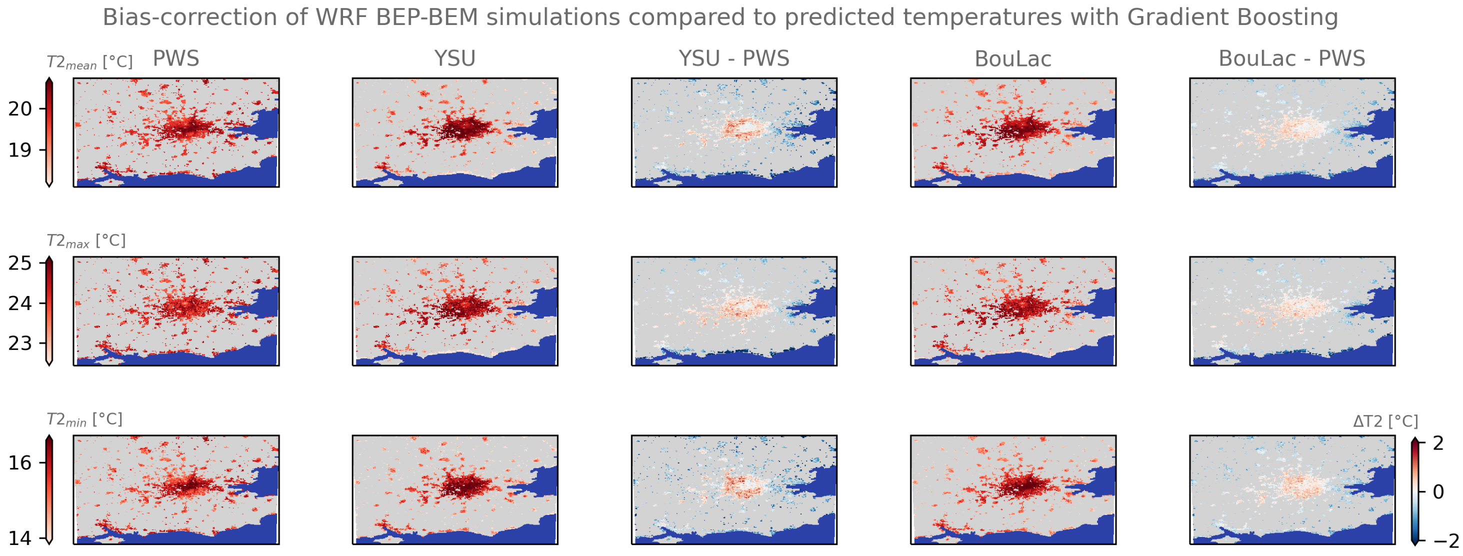


Fig. B11. Same as figure 6, but for gradient boosting regression.

References

Bassett, R., P. Young, G. Blair, F. Samreen, and W. Simm, 2020: A large ensemble approach to quantifying internal model variability within the wrf numerical model. *Journal of Geophysical Research: Atmospheres*, **125 (7)**, e2019JD031 286.

Benjamin, K., Z. Luo, and X. Wang, 2021: Crowdsourcing urban air temperature data for estimating urban heat island and building heating/cooling load in london. *Energies*, **14 (16)**, 5208.

Bougeault, P., and P. Lacarrere, 1989: Parameterization of orography-induced turbulence in a mesobeta–scale model. *Monthly weather review*, **117 (8)**, 1872–1890.

Brisson, E., M. Demuzere, and N. Van Lipzig, 2015: Modelling strategies for performing convection-permitting climate simulations. *Meteorologische Zeitschrift*, **25 (2)**, 149–163.

Broadbent, A. M., J. Declet-Barreto, E. S. Krayenhoff, S. L. Harlan, and M. Georgescu, 2022: Targeted implementation of cool roofs for equitable urban adaptation to extreme heat. *Science of the Total Environment*, **811**, 151 326.

Brousse, O., A. Martilli, M. Foley, G. Mills, and B. Bechtel, 2016: Wudapt, an efficient land use producing data tool for mesoscale models? integration of urban lcz in wrf over madrid. *Urban Climate*, **17**, 116–134.

Brousse, O., C. Simpson, N. Walker, D. Fenner, F. Meier, J. Taylor, and C. Heaviside, 2022: Evidence of horizontal urban heat advection in london using six years of data from a citizen weather station network. *Environmental Research Letters*, **17 (4)**, 044 041.

Chapman, L., C. Bell, and S. Bell, 2017: Can the crowdsourcing data paradigm take atmospheric science to a new level? a case study of the urban heat island of london quantified using netatmo weather stations. *International Journal of Climatology*, **37 (9)**, 3597–3605.

Ching, J., and Coauthors, 2018: Wudapt: An urban weather, climate, and environmental modeling infrastructure for the anthropocene. *Bulletin of the American Meteorological Society*, **99 (9)**, 1907–1924.

Demuzere, M., D. Argüeso, A. Zonato, and J. Kittner, 2021: W2w: A python package that injects wudapt's local climate zone information in wrf (version v0.1.1). Retrieved online, https://pypi.org/project/w2w/.

Demuzere, M., B. Bechtel, A. Middel, and G. Mills, 2019: Mapping europe into local climate zones. *PloS one*, **14 (4)**, e0214 474.

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

Demuzere, M., and Coauthors, 2017: Impact of urban canopy models and external parameters on the modelled urban energy balance in a tropical city. *Quarterly Journal of the Royal Meteorological Society*, **143 (704)**, 1581–1596.

Dudhia, J., 1989: Numerical study of convection observed during the winter monsoon experiment using a mesoscale two-dimensional model. *Journal of Atmospheric Sciences*, **46 (20)**, 3077–3107.

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

Fenner, D., A. Holtmann, F. Meier, I. Langer, and D. Scherer, 2019: Contrasting changes of urban heat island intensity during hot weather episodes. *Environmental Research Letters*, **14 (12)**, 124 013.

Fenner, D., F. Meier, B. Bechtel, M. Otto, and D. Scherer, 2017: Intra and inter 'local climate zone'variability of air temperature as observed by crowdsourced citizen weather stations in berlin, germany. *10.14279/depositonce-10378.*

Georganos, S., and Coauthors, 2021: Geographical random forests: a spatial extension of the random forest algorithm to address spatial heterogeneity in remote sensing and population modelling. *Geocarto International*, **36 (2)**, 121–136.

Grassmann, T., A. Napoly, F. Meier, and D. Fenner, 2018: Quality control for crowdsourced data from cws.

Grimmond, C. S. B., and Coauthors, 2011: Initial results from phase 2 of the international urban energy balance model comparison. *International Journal of Climatology*, **31 (2)**, 244–272.

Gutiérrez, E., J. E. González, A. Martilli, R. Bornstein, and M. Arend, 2015: Simulations of a heat-wave event in new york city using a multilayer urban parameterization. *Journal of Applied Meteorology and Climatology*, **54 (2)**, 283–301.

Hammerberg, K., O. Brousse, A. Martilli, and A. Mahdavi, 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.

Heaviside, C., X.-M. Cai, and S. Vardoulakis, 2015: The effects of horizontal advection on the urban heat island in birmingham and the west midlands, united kingdom during a heatwave. *Quarterly Journal of the Royal Meteorological Society*, **141 (689)**, 1429–1441.

Hendricks, E. A., J. C. Knievel, and Y. Wang, 2020: Addition of multilayer urban canopy models to a nonlocal planetary boundary layer parameterization and evaluation using ideal and real cases. *Journal of Applied Meteorology and Climatology*, **59 (8)**, 1369–1392.

Hollis, D., M. McCarthy, M. Kendon, T. Legg, and I. Simpson, 2019: Haduk-grid—a new uk dataset of gridded climate observations. *Geoscience Data Journal*, **6 (2)**, 151–159.

Hong, S.-Y., J. Dudhia, and S.-H. Chen, 2004: A revised approach to ice microphysical processes for the bulk parameterization of clouds and precipitation. *Monthly weather review*, **132 (1)**, 103–120.

Hong, S.-Y., and S.-W. Kim, 2008: Stable boundary layer mixing in a vertical diffusion scheme. *18th Symposium on Boundary Layers and Turbulence B*, Vol. 16, 325.

Hong, S.-Y., Y. Noh, and J. Dudhia, 2006: A new vertical diffusion package with an explicit treatment of entrainment processes. *Monthly weather review*, **134 (9)**, 2318–2341.

Janić, Z. I., 2001: Nonsingular implementation of the mellor-yamada level 2.5 scheme in the ncep meso model.

Janjić, Z. I., 1994: The step-mountain eta coordinate model: Further developments of the convection, viscous sublayer, and turbulence closure schemes. *Monthly weather review*, **122 (5)**, 927–945.

Jiménez, P. A., J. Dudhia, J. F. González-Rouco, J. Navarro, J. P. Montávez, and E. García-Bustamante, 2012: A revised scheme for the wrf surface layer formulation. *Monthly weather review*, **140 (3)**, 898–918.

Kain, J. S., 2004: The kain–fritsch convective parameterization: an update. *Journal of applied meteorology*, **43 (1)**, 170–181.

Lauwaet, D., H. Hooyberghs, B. Maiheu, W. Lefebvre, G. Driesen, S. Van Looy, and K. De Ridder, 2015: Detailed urban heat island projections for cities worldwide: dynamical downscaling cmip5 global climate models. *Climate*, **3 (2)**, 391–415.

Lipson, M., S. Grimmond, and M. Best, 2021: Urban-plumber model evaluation project: initial results. *EGU General Assembly Conference Abstracts*, EGU21–15 230.

Loridan, T., and C. Grimmond, 2012: Multi-site evaluation of an urban land-surface model: Intra-urban heterogeneity, seasonality and parameter complexity requirements. *Quarterly Journal of the Royal Meteorological Society*, **138 (665)**, 1094–1113.

Maraun, D., and M. Widmann, 2018: *Statistical downscaling and bias correction for climate research*. Cambridge University Press.

Martilli, A., A. Clappier, and M. W. Rotach, 2002: An urban surface exchange parameterisation for mesoscale models. *Boundary-layer meteorology*, **104 (2)**, 261–304.

Martilli, A., and Coauthors, 2021: Simulating the meteorology during persistent wintertime thermal inversions over urban areas. the case of madrid. *Atmospheric Research*, **263**, 105 789.

Masson, V., 2000: A physically-based scheme for the urban energy budget in atmospheric models. *Boundary-layer meteorology*, **94 (3)**, 357–397.

McCarthy, M., and Coauthors, 2019: Drivers of the uk summer heatwave of 2018. *Weather*, **74 (11)**, 390–396.

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

Mlawer, E. J., S. J. Taubman, P. D. Brown, M. J. Iacono, and S. A. Clough, 1997: Radiative transfer for inhomogeneous atmospheres: Rrtm, a validated correlated-k model for the longwave. *Journal of Geophysical Research: Atmospheres*, **102 (D14)**, 16 663–16 682.

Mughal, M. O., X.-X. Li, T. Yin, A. Martilli, O. Brousse, M. A. Dissegna, and L. K. Norford, 2019: High-resolution, multilayer modeling of singapore's urban climate incorporating local climate zones. *Journal of Geophysical Research: Atmospheres*, **124 (14)**, 7764–7785.

Muller, C., L. Chapman, S. Johnston, C. Kidd, S. Illingworth, G. Foody, A. Overeem, and R. Leigh, 2015: Crowdsourcing for climate and atmospheric sciences: current status and future potential. *International Journal of Climatology*, **35 (11)**, 3185–3203.

Napoly, A., T. Grassmann, F. Meier, and D. Fenner, 2018: Development and application of a statistically-based quality control for crowdsourced air temperature data. *Frontiers in Earth Science*, **6**, 118.

Nazarian, N., and Coauthors, 2022: Integrated assessment of urban overheating impacts on human life. *Earth's Future*.

Niu, G.-Y., and Coauthors, 2011: The community noah land surface model with multiparameterization options (noah-mp): 1. model description and evaluation with local-scale measurements. *Journal of Geophysical Research: Atmospheres*, **116 (D12)**.

Oke, T. R., G. Mills, A. Christen, and J. A. Voogt, 2017: *Urban climates*. Cambridge University Press.

Oleson, K., G. Anderson, B. Jones, S. McGinnis, and B. Sanderson, 2018: Avoided climate impacts of urban and rural heat and cold waves over the us using large climate model ensembles for rcp8. 5 and rcp4. 5. *Climatic change*, **146 (3)**, 377–392.

Pedregosa, F., and Coauthors, 2011: Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, **12**, 2825–2830.

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

Salamanca, F., A. Krpo, A. Martilli, and A. Clappier, 2010: A new building energy model coupled with an urban canopy parameterization for urban climate simulations—part i. formulation, verification, and sensitivity analysis of the model. *Theoretical and applied climatology*, **99 (3)**, 331–344.

Salamanca, F., and A. Martilli, 2010: A new building energy model coupled with an urban canopy parameterization for urban climate simulations—part ii. validation with one dimension off-line simulations. *Theoretical and Applied Climatology*, **99 (3)**, 345–356.

Salamanca, F., A. Martilli, M. Tewari, and F. Chen, 2011: A study of the urban boundary layer using different urban parameterizations and high-resolution urban canopy parameters with wrf. *Journal of Applied Meteorology and Climatology*, **50 (5)**, 1107–1128.

Salamanca, F., A. Martilli, and C. Yagüe, 2012: A numerical study of the urban heat island over madrid during the desirex (2008) campaign with wrf and an evaluation of simple mitigation strategies. *International Journal of Climatology*, **32 (15)**, 2372–2386.

Sgoff, C., W. Acevedo, Z. Paschalidi, S. Ulbrich, E. Bauernschubert, T. Kratzsch, and R. Potthast, 2022: Assimilation of crowd-sourced surface observations over germany in a regional weather prediction system. *Quarterly Journal of the Royal Meteorological Society*.

Stewart, I. D., and T. R. Oke, 2012: Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society*, **93 (12)**, 1879–1900.

Stewart, I. D., T. R. Oke, and E. S. Krayenhoff, 2014: Evaluation of the ‘local climate zone’scheme using temperature observations and model simulations. *International journal of climatology*, **34 (4)**, 1062–1080.

Sunter, M., 2021: Midas data user guide for uk land observations, v20210705.

Tewari, M., F. Salamanca, A. Martilli, L. Treinish, and A. Mahalov, 2017: Impacts of projected urban expansion and global warming on cooling energy demand over a semiarid region. *Atmospheric Science Letters*, **18 (11)**, 419–426.

UKMO, 2021: Midas open: Uk hourly weather observation data, v202107. centre for environmental data analysis, 08 september 2021. Data retrieved online, doi:10.5285/3bd7221d4844435dad2fa030f26ab5fd.

Varentsov, M., D. Fenner, F. Meier, T. Samsonov, and M. Demuzere, 2021: Quantifying local and mesoscale drivers of the urban heat island of moscow with reference and crowdsourced observations. *Frontiers in Environmental Science*, 543.

Venter, Z. S., O. Brousse, I. Esau, and F. Meier, 2020: Hyperlocal mapping of urban air temperature using remote sensing and crowdsourced weather data. *Remote Sensing of Environment*, **242**, 111 791.

Venter, Z. S., T. Chakraborty, and X. Lee, 2021: Crowdsourced air temperatures contrast satellite measures of the urban heat island and its mechanisms. *Science Advances*, **7 (22)**, eabb9569.

Virtanen, P., and Coauthors, 2020: Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, **17 (3)**, 261–272.

Wang, J., and X.-M. Hu, 2021: Evaluating the performance of wrf urban schemes and pbl schemes over dallas–fort worth during a dry summer and a wet summer. *Journal of Applied Meteorology and Climatology*, **60 (6)**, 779–798.

Wouters, H., M. Demuzere, U. Blahak, K. Fortuniak, B. Maiheu, J. Camps, D. Tielemans, and N. P. van Lipzig, 2016: The efficient urban canopy dependency parametrization (sury) v1. 0 for atmospheric modelling: description and application with the cosmo-clm model for a belgian summer. *Geoscientific Model Development*, **9 (9)**, 3027–3054.

Wouters, H., and Coauthors, 2017: Heat stress increase under climate change twice as large in cities as in rural areas: A study for a densely populated midlatitude maritime region. *Geophysical Research Letters*, **44 (17)**, 8997–9007.

Yang, J., and E. Bou-Zeid, 2019: Scale dependence of the benefits and efficiency of green and cool roofs. *Landscape and urban planning*, **185**, 127–140.

Yang, Z.-L., and Coauthors, 2011: The community noah land surface model with multiparameterization options (noah-mp): 2. evaluation over global river basins. *Journal of Geophysical Research: Atmospheres*, **116 (D12)**.

Zängl, G., D. Reinert, P. Rípodas, and M. Baldauf, 2015: The icon (icosahedral non-hydrostatic) modelling framework of dwd and mpi-m: Description of the non-hydrostatic dynamical core. *Quarterly Journal of the Royal Meteorological Society*, **141 (687)**, 563–579.

Zonato, A., A. Martilli, S. Di Sabatino, D. Zardi, and L. Giovannini, 2020: Evaluating the performance of a novel wudapt averaging technique to define urban morphology with mesoscale models. *Urban Climate*, **31**, 100 584.

Zonato, A., A. Martilli, P. A. Jimenez, J. Dudhia, D. Zardi, and L. Giovannini, 2022: A new k–ε turbulence parameterization for mesoscale meteorological models. *Monthly Weather Review*.

Zumwald, M., B. Knüsel, D. N. Bresch, and R. Knutti, 2021: Mapping urban temperature using crowd-sensing data and machine learning. *Urban Climate*, **35**, 100 739.