Received: 28 February 2023 | Revised: 8 September 2023 | Accepted: 27 September 2023

DOI: 10.1002/qj.4589

Quarterly Journal of the Royal Meteorological Society RMetS

RESEARCH ARTICLE

Evaluation of 30 urban land surface models in the Urban-PLUMBER project: Phase 1 results

Mathew J. Lipson[1,2] | **Sue Grimmond**[3] | **Martin Best**[4] | **Gab Abramowitz**[5] | **Andrew Coutts**[6] | **Nigel Tapper**[6] | **Jong-Jin Baik**[7] | **Meiring Beyers**[8] | **Lewis Blunn**[9] | **Souhail Boussetta**[10] | **Elie Bou-Zeid**[11] | **Martin G. De Kauwe**[12] | **Cécile de Munck**[13] | **Matthias Demuzere**[14] | **Simone Fatichi**[15] | **Krzysztof Fortuniak**[16] | **Beom-Soon Han**[17] | **Margaret A. Hendry**[4] | **Yukihiro Kikegawa**[18] | **Hiroaki Kondo**[19,20] | **Doo-Il Lee**[21] | **Sang-Hyun Lee**[21] | **Aude Lemonsu**[13] | **Tiago Machado**[13] | **Gabriele Manoli**[22] | **Alberto Martilli**[23] | **Valéry Masson**[13] | **Joe McNorton**[10] | **Naika Meili**[15] | **David Meyer**[3,24] | **Kerry A. Nice**[25] | **Keith W. Oleson**[26] | **Seung-Bu Park**[27] | **Michael Roth**[8] | **Robert Schoetter**[13] | **Andrés Simón-Moral**[28] | **Gert-Jan Steeneveld**[29] | **Ting Sun**[30] | **Yuya Takane**[19] | **Marcus Thatcher**[31] | **Aristofanis Tsiringakis**[32] | **Mikhail Varentsov**[33,34] | **Chenghao Wang**[35,36] | **Zhi-Hua Wang**[37] | **Andy J. Pitman**[5]

Correspondence
Mathew J. Lipson, Australian Research Council Centre of Excellence for Climate System Science, Climate Change Research Centre, Level 4, Mathews Building, UNSW Sydney, New South Wales, 2052, Australia; Bureau of Meteorology, Sydney, NSW, Australia.
Email: mathew.lipson@bom.gov.au

Funding information
Army Research Office, Grant/Award Number: W911NF2010216; Australian Research Council, Grant/Award Numbers: CE110001028, CE170100023; Bureau of Meteorology, Australian Government; Deutsche Forschungsgemeinschaft, Grant/Award Number: 437467569; H2020 European Research Council, Grant/Award Number: 855005; Japan Society for the Promotion of Science, Grant/Award Number: 20KK0096; Met Office; National Computational Infrastructure; National Health and Medical Research Council, Grant/Award

Abstract

Accurately predicting weather and climate in cities is critical for safeguarding human health and strengthening urban resilience. Multimodel evaluations can lead to model improvements; however, there have been no major intercomparisons of urban-focussed land surface models in over a decade. Here, in Phase 1 of the Urban-PLUMBER project, we evaluate the ability of 30 land surface models to simulate surface energy fluxes critical to atmospheric meteorological and air quality simulations. We establish minimum and upper performance expectations for participating models using simple information-limited models as benchmarks. Compared with the last major model intercomparison at the same site, we find broad improvement in the current cohort's predictions of short-wave radiation, sensible and latent heat fluxes, but little or no improvement in long-wave radiation and momentum fluxes. Models with a simple urban representation (e.g., 'slab' schemes) generally perform well, particularly when combined with sophisticated hydrological/vegetation models. Some mid-complexity models (e.g., 'canyon' schemes) also perform well, indicating efforts to integrate vegetation and hydrology processes have paid dividends. The most complex models that resolve three-dimensional interactions between

For affiliations refer to page 150

Number: 1194959; National Research Foundation of Korea, Grant/Award Number: 2021R1A2C1007044; National Science Foundation, Grant/Award Numbers: 1852977, AGS 2128345; National University of Singapore, Grant/Award Number: 22-3637-A0001; Natural Environment Research Council, Grant/Award Number: NE/W010003/1; Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Number: 864.14.007; Nuclear Safety and Security Commission, Grant/Award Number: 2105036; Russian Science Foundation, Grant/Award Number: 21-17-00249; UK Research and Innovation, Grant/Award Numbers: NE/P018637/1, NE/P018637/2; University of New South Wales; University of Reading; Korea Meteorological Administration Research and Development Program, Grant/Award Numbers: KMI(KMI2021-03512)

buildings in general did not perform as well as other categories. However, these models also tended to have the simplest representations of hydrology and vegetation. Models without any urban representation (i.e., vegetation-only land surface models) performed poorly for latent heat fluxes, and reasonably for other energy fluxes at this suburban site. Our analysis identified widespread human errors in initial submissions that substantially affected model performances. Although significant efforts are applied to correct these errors, we conclude that human factors are likely to influence results in this (or any) model intercomparison, particularly where participating scientists have varying experience and first languages. These initial results are for one suburban site, and future phases of Urban-PLUMBER will evaluate models across 20 sites in different urban and regional climate zones.

KEYWORDS

benchmark, energy balance, intercomparison, model evaluation, urban climate, urban meteorology

1 | INTRODUCTION

Over a decade has passed since the first International Urban Land Surface Model Comparison Project (PILPS-Urban) evaluated 32 models at two urban sites (Grimmond *et al.*, 2010, 2011). Since then, new urban models have been developed, existing models have increased capabilities, and a new generation of modellers are using them. Expectations that urban schemes be integrated within weather and climate models have also grown: simulations are undertaken at finer spatial scales, and the wider modelling community recognises the importance of simulating meteorological conditions within cities (Masson *et al.*, 2020; Sharma *et al.*, 2021). Therefore, it is timely to undertake a new evaluation of land surface models used in meteorological simulations over urban areas.

This project, Urban-PLUMBER, focusses on the local-scale (order 0.1–5 km) energy exchange between the urban land surface and the atmosphere. The last intercomparison with a similar focus, PILPS-Urban (Grimmond *et al.*, 2010, 2011) established that knowledge of an urban site's surface cover fractions significantly improved model performance. 'Urban' models can include impervious surfaces (e.g., buildings, roads) and pervious surfaces (e.g., vegetation, bare earth), but not all urban models include both. In PILPS-Urban, models that neglected vegetation or porous ground performed poorly in latent, sensible, and radiant heat fluxes. This may have been expected at a suburban site with about 40% vegetation fraction (Grimmond *et al.*, 2011); however, their performances were also poorer at an urban site nearly devoid of vegetation (Grimmond *et al.*, 2010). PILPS-Urban concluded that models with simpler urban geometry (i.e., with fewer parameters describing built up areas) generally performed better than more complex models, as simpler models were better able to use provided site information. Further analysis of the suburban-site results (Best and Grimmond, 2015) concluded the dominant physical processes that urban models should capture, by importance, are: (a) bulk surface albedo during the day; (b) trapping of long-wave radiation between urban structures at night; and (c) evapotranspiration over diurnal and seasonal timescales.

Urban-PLUMBER builds on PILPS-Urban, which in turn drew on the methods of PILPS (Project for the Intercomparison of Land Surface Parameterisation Schemes). Since the 1990s, PILPS projects have undertaken land surface model evaluation and comparison (Henderson-Sellers *et al.*, 1996, 1995; Slater *et al.*, 2001; Bowling *et al.*, 2003). A coordinating group defines the project framework (the protocol) and provides participating modelling groups with both meteorological data to drive land surface models and surface characteristics parameters to configure models. After running a model on their computers, participants submit their outputs to coordinators. Coordinators analyse outputs and communicate results. More generally, model intercomparison projects (MIPs) have been undertaken across all Earth system spheres, and have become a foundational element of climate science (Eyring *et al.*, 2016). Together with PILPS-Urban, two additional MIPs have been influential in the design of the current project.

PLUMBER (Protocol for the Analysis of Land Surface Models Benchmarking Evaluation Project) (Best *et al.*, 2015) demonstrated the benefit of using benchmarks to set the performance expectations for models. In

traditional modelcomparison, models are ranked by various error metrics for select observed outputs. Although this helps identify outlying model performances it does not help determine whether the cohort overall is performing well or poorly. Furthermore, it may lead to subjective assessment of models being (un)fit for purpose and misdirecting subsequent model development priorities (Best *et al.*, 2015). The benchmarks used in PLUMBER were simple empirical and physically-based models with far fewer inputs than the participating land surface models. Comparing models with benchmarks indicates the strengths and weaknesses of the cohort and hence areas for future development. Unsettlingly, PLUMBER found very simple empirical models such as a linear regression driven by short-wave radiation observed at other sites (i.e., trained out-of-sample) outperformed all participating land surface models across a suite of standard metrics when predicting sensible heat fluxes over 20 sites. The authors of the PLUMBER project concluded that complex and computationally expensive land surface models were not effectively using the information available in the forcing data when determining surface–atmosphere turbulent fluxes, arguing this challenged broadly accepted concepts used to model the surface energy balance. In Urban-PLUMBER we adopt a similar benchmarking approach but apply it for the first time in an urban setting.

ESM-SnowMIP (Earth System Model–Snow Model Intercomparison Project) (Menard *et al.*, 2021) found widespread human errors affecting model performance but, unlike some earlier comparisons, it encouraged resubmissions where initial results showed unexpected behaviour. As such, most modellers resubmitted their results when errors were identified. Errors included incorrectly configuring model start times, using input data from the wrong sites, incorrectly formatted model outputs (variable name or sign), and hardcoded bugs (i.e., coding errors in model parameterisation). In the same way, Urban-PLUMBER aims to reduce human errors via an initial assessment and resubmission process to better focus on intended model functionality.

The Urban-PLUMBER project involves 30 models (Table 1, Appendix A1), 20 urban sites (Lipson *et al.*, 2022a), 50 site-years of meteorological observations, 200 site-years of synthetic data, and 55 model output variables. Here, in Phase 1 of the project, we focus on evaluating model performance of five observed surface–atmosphere fluxes at one suburban site (Preston, Melbourne, Australia) over 16 months. The same site and observational data were used in PILPS-Urban (Grimmond *et al.*, 2011), allowing direct comparison with those results; hence, our objectives here are to (a) evaluate land surface model performance in an urban setting using a benchmarking methodology; and (b) assess how the current cohort of models compare to earlier participants of PILPS-Urban.

2 | METHODS

2.1 | Overview of modelling approaches

Many urban land surface models exist to parameterise urban surface–atmosphere exchanges (Grimmond *et al.*, 2009; Garuma, 2018; Nazarian *et al.*, 2023) and are developed for various purposes including to predict lower boundary conditions for weather, climate or air quality simulations; forecast environmental conditions within the urban canopy (e.g., between buildings at pedestrian level); test interventions intended to improve these conditions; and predict anthropogenic feedbacks relating to energy and water use or thermal comfort.

Although there is effectively a continuum of models with different levels of complexity for different physical processes (Figures 1 and 2, Appendix A1), here we broadly classify models into one of five cohorts (Figure 2) based on the representation of urban impervious surfaces (buildings, roads etc):

- *Non-urban schemes (participants in cohort n = 2)*: Most global and some regional weather and climate models lack an explicit urban scheme (Best, 2006; Oleson *et al.*, 2018; Daniel *et al.*, 2019; Zhao *et al.*, 2021). Rather they simulate these areas using bare earth, rock or vegetation. Including models in this class helps determine the importance of using an urban scheme at a suburban site.
- *One-tile (slab) schemes (n = 5)*: These treat built areas as a homogenous flat surface with parameters modified to represent the bulk influence of all urban elements. Some one-tile urban schemes represent built urban elements only (buildings, paving, roads etc), while others include the effects of vegetation and other surface types (water, bare soil, etc). Therefore, optimal effective bulk surface parameters are model-, site- and output-specific (Salamanca *et al.*, 2009). Methods to estimate effective surface parameters include tuning to appropriately scaled observations (Best *et al.*, 2006), from more detailed models (Martilli *et al.*, 2015), or from more detailed input data (Wouters *et al.*, 2016).
- *Two-tile schemes (n = 5)*: These resolve two urban surface facets (e.g., roofs and 'street canyons') with different thermal and radiative properties, and therefore different surface energy balances. Best *et al.* (2006) suggested two-tile schemes provide benefit because one-tile heat capacity values could not be selected which provide both the correct amplitude and phase

TABLE 1 Participating models. Table 2 and Appendix A1 provide further details for each model

ID	Submission name	Urban land surface model	Vegetation land surface model (if distinct from urban model)
01	ASLUMv2.0	Arizona State University single-layer urban canopy model v2.0	(Integrated vegetation)
02	ASLUMv3.1	Arizona State University single-layer urban canopy model v3.1	(Integrated vegetation)
03	BEPCOL	Building effect parameterisation – column model	Bare soil model based on regional atmospheric modelling system (RAMS)
04	CABLE	–	Community atmosphere–biosphere land exchange model
05	CHTESSEL	–	Carbon hydrology tiled ECMWF scheme for surface exchanges over land (CHTESSEL)
06	CHTESSEL_U	Urban scheme from CHTESSEL	Tiled ECMWF scheme for surface exchanges over land (CHTESSEL)
07	CLMU5	Community land model urban	(Integrated vegetation)
08	CM	Canopy model	(Integrated vegetation)
09	CM-BEM	Canopy model – building energy model	(Integrated vegetation)
10	JULES_1T	One-tile urban scheme from JULES	Joint UK land environment simulator (JULES)
11	JULES_2T	Two-tile urban scheme from JULES	Joint UK land environment simulator (JULES)
12	JULES_MORUSES	Met Office Reading urban exchange scheme	Joint UK land environment simulator (JULES)
13	K-UCMv1	Klimaat urban canopy model	(Integrated vegetation)
14	Lodz-SUEB	Lodz SUrface energy balance	(Integrated vegetation)
15	Manabe_1T	One-tile urban scheme from JULES	Manabe bucket
16	Manabe_2T	Two-tile urban scheme from JULES	Manabe bucket
17	MUSE	Microscale urban surface energy model	Bowen ratio method
18	NOAH-SLAB	Slab urban scheme from Noah-LSM	Noah land surface model (Noah-LSM)
19	NOAH-SLUCM	Single-layer urban canopy model (SLUCM)	Noah land surface model (Noah-LSM)
20	SNUUCM	Seoul National University urban canopy model	Noah land surface model (Noah-LSM)
21	SUEWS	Surface urban energy and water balance scheme	(Integrated vegetation)
22	TARGET	The Air Temperature Response to Green/blue-infrastructure Evaluation Tool (TARGET)	(Integrated vegetation)
23	TEB-CNRM	Town energy balance (TEB) with road canyon hypothesis for radiation	ISBA (included in SURFEX)
24	TEB-READING	Town energy balance (TEB) with road canyon hypothesis for radiation	Simple partitioning using fixed Bowen ratio and albedo
25	TEB-SPARTCS	Town energy balance with SPARTACUS-urban for radiative exchanges	ISBA (included in SURFEX)
26	TERRA_4.11	TERRA_URB	TERRA (stand-alone version)
27	UCLEM	Urban Climate and energy model (UCLEM)	(Integrated vegetation)
28	UT&C	Urban Tethys-Chloris (UT&C)	(Integrated vegetation)
29	VTUF-3D	Vegetated temperatures of urban facets (VTUF)	MAESPA
30	VUCM	Vegetated urban canopy model (VUCM)	(Integrated vegetation)

Note: Section 2.1 provides an overview of urban modelling approaches and references.

TABLE 2 Participating model information

ID	Submission name	Version(s)	Scheme(s)	Scale(s)	Primary purpose(s)	Participating author(s)
01	ASLUMv2.0	v2.0	U	L/R	CF/TA/TC/WS/SEB	Wang, Wang
02	ASLUMv3.1	v3.1	U	L/R	CF/TA/TC/WS/SEB	Wang, Wang
03	BEPCOL	v1	U/V	L	TA/SEB	Simón-Moral, Martilli
04	CABLE	CABLE trunk r7025	V	G	CF	De Kauwe
05	CHTESSEL	CHTESSEL-IFS-CY47R1	V	G	CF	McNorton, Boussetta
06	CHTESSEL_U	CHTESSEL-IFS-CY47R1_URBAN	U/V	G	CF	McNorton, Boussetta
07	CLMU5	Release-clm5.0.34	U	R/G	CF	Oleson
08	CM	CM v2021	U	R/G	TA/SEB	Takane, Kondo
09	CM-BEM	CM-BEM v2021	U	R/G	TA/TC/SEB/E	Takane, Kikegawa
10	JULES_1T	GL9	U	R/G	CF/O	Best
11	JULES_2T	GL9	U	R/G	CF	Best
12	JULES_MORUSES	GL9	U	R/G	CF/O	Hendry, Best
13	K-UCMv1	v1	U/V	L	TA/TC/SEB	Beyers, Roth
14	Lodz-SUEB	v3	U	L	SEB	Fortuniak
15	Manabe_1T	GL9	U/V	L	SEB/BM	Best
16	Manabe_2T	GL9	U/V	L	SEB/BM	Best
17	MUSE	V1.0	U	M/L	CF/TC/SEB	Lee, Lee
18	NOAH-SLAB	Noah-LSM v3.4.1	U/V	L	CF	Steeneveld, Tsiringakis
19	NOAH-SLUCM	Noah-LSM v3.4.1	U/V	L/R	CF/TA/SEB	Tsiringakis, Steeneveld
20	SNUUCM	SNUUCM+Noah-LSM v1.0	U/V	L/R	CF/AQ	Park, Baik
21	SUEWS	SUEWS v2020a	U	L	TA/TC/WS/H/SEB	Sun, Blunn
22	TARGET	TARGET-Java v1.1	U/V	L	TA/TC/WS	Nice
23	TEB-CNRM	SURFEX v9	U	R/G	CF/TA/TC/WS/H/SEB/O	Machado, de Munck, Schoetter, Masson, Lemonsu
24	TEB-READING	TEB v4.1.0	U/V	R	CF/TA/SEB	Meyer
25	TEB-SPARTCS	SURFEX v9	U	R/G	CF/TA/TC/WS/H/SEB	Machado, de Munck, Schoetter, Masson, Lemonsu
26	TERRA_4.11	v4.11	U/V	L/R	CF/AQ/TC/WS/SEB/O	Demuzere, Varentsov
27	UCLEM	CCAM r4909	U	G	CF/E	Thatcher, Lipson
28	UT&C	v1.0	U/V	L/R	TA/TC/WS/H/SEB	Meili, Fatichi, Manoli, Bou-Zeid
29	VTUF-3D	Java v1.0	U	M	TA/TC/WS/SEB	Nice
30	VUCM	V1.0	U	M/L	CF/AQ/TA/TC/WS/H	Lee, Han

Note: Model may include an U: urban and/or V: vegetation land surface scheme, scale developed for M: micro, L: local, R: regional, G: global; and intended purpose to simulate CF: climate and weather forecasting, AQ: air quality, TA: temperature of air in canopy, TC: thermal comfort, WS: water-sensitive urban design, E: energy consumption analysis, H: hydrological analysis, SEB: surface energy balance, O: operational model for numerical weather prediction, or as a BM: benchmark for this study.

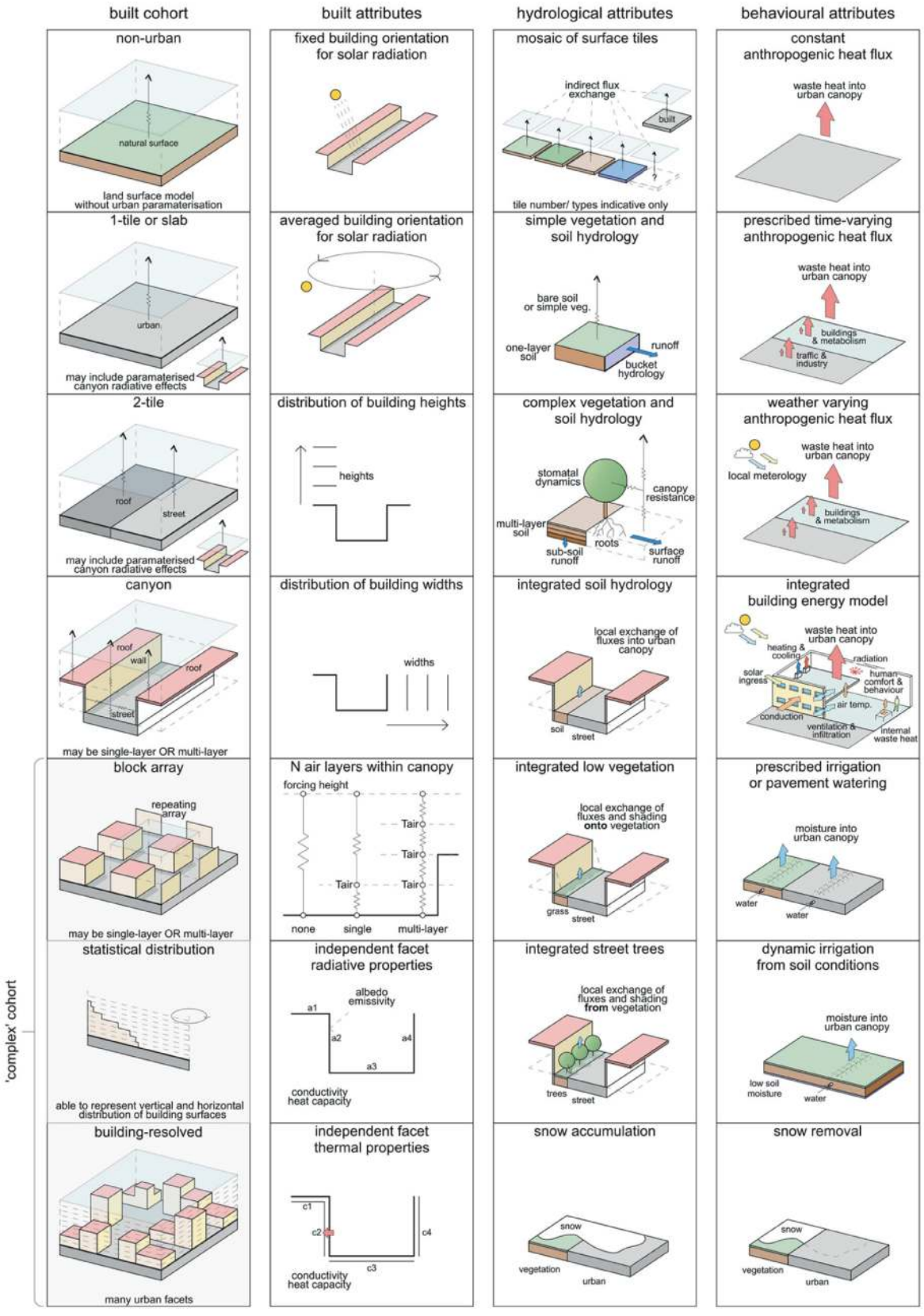

FIGURE 1 Model schematics of the main built, hydrological and behavioural attributes for participating models. Here models are categorised into five cohorts (left column) based on the geometric representation of buildings, with built, hydrological and behavioural attributes used to refine a 'total complexity' (Figure 2). Block array, statistical distribution and building-resolved models are grouped together into a 'complex' cohort in later analysis.

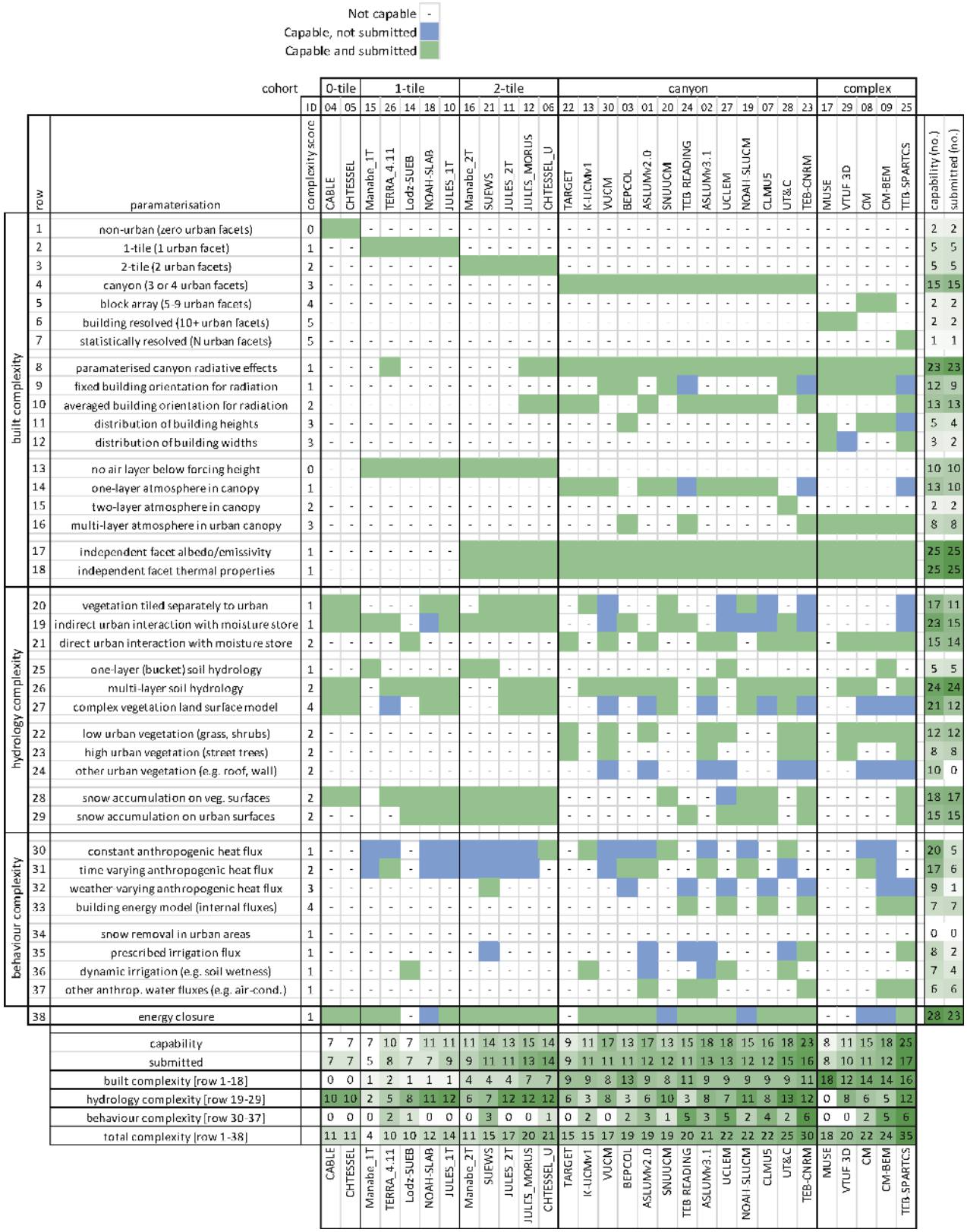

FIGURE 2 Participating model capabilities. Grouped into cohorts (non-urban, one-tile, two-tile, canyon, complex: defined by approach to the built part of the urban area) and sorted from lower to higher 'total complexity' as calculated by assessing the included built, hydrological and behavioural attributes of submissions (green cells). Blue cells indicate a model is capable of representing the process but was not used in this submission. Frequency of approaches are indicated in right column. The 'complexity score' for each process is subjective. It is intended to be indicative only, helping to distinguish models within cohorts.

for observed sensible heat fluxes. While most two-tile schemes have surface parameters constant throughout a simulation, some parameterise the radiative and thermal effects of canyons from sun angle and morphology (e.g., MORUSES – Porson *et al.*, 2010; CHTESSEL_U – McNorton *et al.*, 2021). In this project these are classified as two-tile rather than canyon schemes, as they resolve two surface energy balances only.

- *Canyon schemes (n = 13)*: These resolve the energy balance for roof, wall and ground surfaces separately (Masson, 2000). Radiation reflection and trapping are simulated in two dimensions with an infinite canyon assumption. The details of canyon schemes vary widely, with single or multiple atmosphere layers, sub-facets (e.g., multifaceted walls), fixed or averaged building orientation, independent facet thermal and radiative properties, constant or distributed building heights, and those that include pervious ground, low vegetation and/or street trees between buildings (Figures 1 and 2).
- *More complex schemes (n = 5)*: These resolve three-dimensional interactions between urban facets using a variety of approaches. Repeated cuboids allow for two perpendicular streets while retaining some of the computational efficiency of a canyon approach (Kanda *et al.*, 2005). Statistical distributions can characterise realistic urban environments and have been used to determine three-dimensional radiative interactions between buildings and urban vegetation similar to three-dimensional radiative interactions between clouds (Hogan, 2019a, 2019b). This allows complex urban environments to be simulated in a computationally efficient manner (Stretton *et al.*, 2022). Building- and tree-resolving models represent three-dimensional interactions more explicitly, allowing microclimate conditions to be resolved, but at a larger computational cost.

Models can be further distinguished by how or if hydrological and anthropogenic processes are addressed, again with a large variety of approaches (Figure 1). How models represent built, hydrological and anthropogenic processes are used here to obtain a measure of each model's 'total complexity' (Figure 2). Participating model's parameterisations are individually summarised in Appendix A1.

2.2 | Experiment design and data

2.2.1 | Site description

Simulations are undertaken for the Preston area in Melbourne, Australia (AU-Preston; Lipson *et al.*, 2022a), the same site used in PILPS-Urban Phase 2 (Grimmond *et al.*, 2011). The site area includes 1–2-storey detached residential buildings, some row-style 1–2-storey commercial buildings, and substantial tree and lawn cover (Figure 3). The neighbourhood is classed as an open low-rise (LCZ6) Local Climate Zone (Stewart and Oke, 2012; Demuzere *et al.*, 2022). The region is classified as having a temperate oceanic climate (CfB) under the Köppen–Geiger system (Beck *et al.*, 2018).

The site parameter values (Table 3) provided to participants are drawn from publications (Coutts, 2006; Coutts *et al.*, 2007a, 2007b; Grimmond *et al.*, 2011; Nice *et al.*, 2018) or, when unavailable, estimated from high-resolution global datasets (e.g., OpenLandMap soil datasets; Hengl, 2018a, 2018b, 2018c).

2.2.2 | Observational and forcing data

Observations for the AU-Preston site were gathered using sensors mounted on a telecommunication tower 40 m above ground to measure local-scale conditions (i.e., rather than microscale; Coutts *et al.*, 2007a). Measurement height is 6.25 times mean building height (Table 3) and is thus assumed to be within the constant flux layer and inertial sub-layer. Raw data were obtained over 474.4 days (12 August 2003 to 28 November 2004) at high frequency (1–10 Hz), which are then quality-controlled and averaged to 30-min with period-ending timestamps. Quality control removes periods unsuitable for eddy covariance observations (e.g., strongly stable conditions or periods subject to flow interference), along with significant outliers and unphysical values (Coutts *et al.*, 2007a, 2007b; Lipson *et al.*, 2022a).

The site observations are split into: (a) *forcing data*: provided to participants to drive models; and (b) *analysis data*: withheld from participants and used to evaluate model performances (Table 4). Analysis data are not gap-filled; models are evaluated against observed data only, and not analysed during periods with gap-filled short-wave down (SW_{down}; except where $SW_{down} = 0$ at night, which is assumed valid). SW_{up} is not analysed at night. After quality control and periods of equipment failure, remaining analysis data are well spread between day and night, and across the four seasons (Table 4). Additional processing description, observational data and plots are included in Lipson *et al.* (2022c).

The forcing dataset is gap-filled since it needs to be continuous for models. Small gaps (≤2 hr) are filled by linearly interpolating from available data. Larger gaps are filled using ERA5 global reanalysis (Hersbach *et al.*, 2020) hourly data on single levels at 0.25° spatial resolution (Hersbach *et al.*, 2018). As gridded data differ from point observations (Martens *et al.*, 2020), and ERA5 does not use a model with urban climate effects

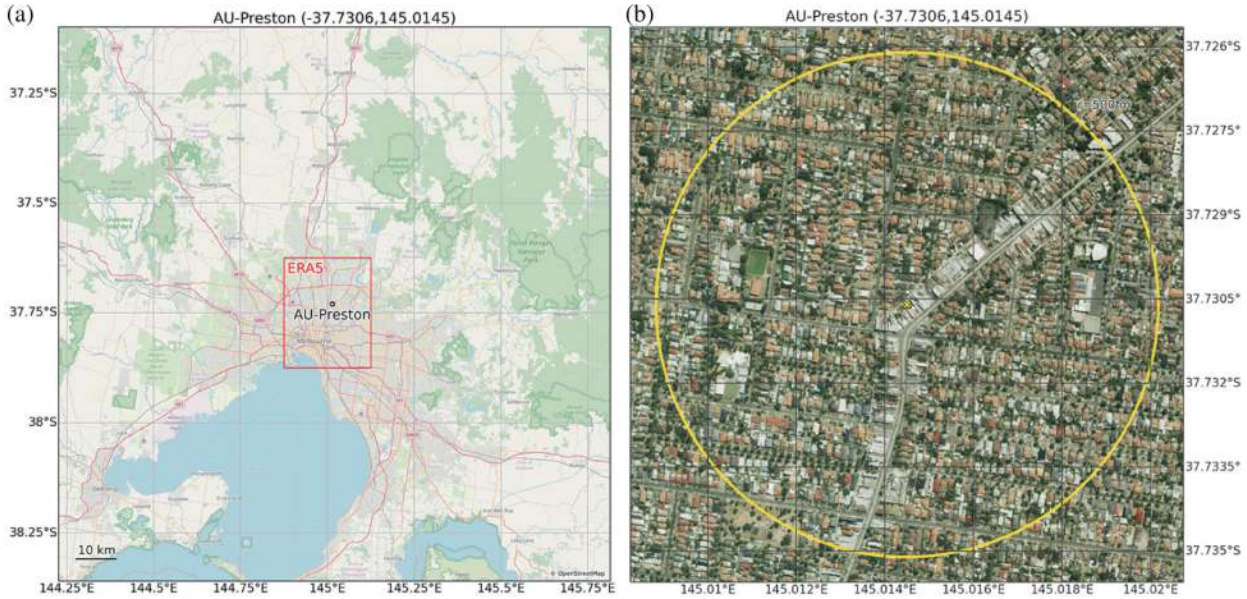

FIGURE 3 Study area, AU-Preston: (a) location within Melbourne, Australia with the extent of the ERA5 (Hersbach *et al.*, 2020) grid cell used for gap-filling observations (red rectangle) (image: © OpenStreetMap contributors); and (b) aerial imagery around the flux tower site (yellow cross with circle of 500 m radius) (image: © State of Victoria [Department of Environment, Land, Water and Planning]).

(McNorton *et al.*, 2021), diurnal and seasonal adjustments are applied to bias-correct ERA5 data using available site observations and nearby rain gauges before gap-filling (Lipson *et al.*, 2022a).

Systematic and random errors are present in any observations used to force and evaluate models. Random errors in flux observations over forested areas generally scale with the magnitude of the flux (Hollinger and Richardson, 2005; Richardson *et al.*, 2006). Flux observations at urban sites have reported random and systematic uncertainties in the same range as observed over vegetated ecosystems (Järvi *et al.*, 2018). At this site, daytime flux errors have been estimated to be up to 10% (Best and Grimmond, 2015). Evaluating models over extended periods reduces the effects of random errors. However, we cannot account for systematic errors if they exist, nor can we assess if surface energy closure is achieved with the available observations.

Annualised rainfall during the analysis period (682 mm) was near the long-term average. However, preceding drought conditions and ongoing restrictions on domestic irrigation led to lower moisture availability and higher Bowen ratios during the study period (Coutts *et al.*, 2007b). Conditions were otherwise reasonably representative of typical local climatology (Lipson *et al.*, 2022c).

2.2.3 | Spin-up strategy

Soil wetness at the beginning of a simulation (the initial conditions) can strongly influence the modelled surface energy balance. Most land surface models require years to reach a hydrological equilibrium when forced by local meteorology (Yang *et al.*, 1995; Best and Grimmond, 2014). As soil states are model-dependent, initial conditions cannot simply be transferred between models nor set to one state across models (Koster *et al.*, 2009). Ideally, each model would reach their own equilibrium during a spin-up period which is not analysed, with 10 years considered generally sufficient across a wide range of land surface models (Best *et al.*, 2015; Best and Grimmond, 2016b).

As model-forcing observations are rarely available to allow such a long spin-up at urban sites, past evaluation strategies include discarding some initial observations as spin-up (Grimmond *et al.*, 2011), repeating a single year of observations several times (Best *et al.*, 2015), using global reanalysis products such as ERA5 (Hersbach *et al.*, 2020) or reanalysis data with bias corrections applied from gridded observations, such as WFDE5 (Cucchi *et al.*, 2020). Using reanalysis for spin-up represents interannual variability prior to the analysis period and allows observations to be used for analysis. However, gridded reanalysis data (with grid spacing of order 30 km or coarser) may be unsatisfactory if local urban effects are not captured. To address this, we use site-bias-corrected ERA5 time series for 10 years prior to analysis (Lipson *et al.*, 2022a). This provides meteorology (precipitation, solar radiation, temperature, wind etc.) over a sufficiently long period for soil states to equilibrate with local conditions prior to the analysis period. Of the 30 participating models, five did not use the full spin-up period (ASLUMv2.0, ASLUMv3.1,

TABLE 3 Site-descriptive metadata

ID	Parameter	Value	Units	Footprint	Source
Baseline experiment parameters (1–9)					
1	Latitude	−37.7306	Degrees_north	Tower	(Coutts *et al.*, 2007a)
2	Longitude	145.0145	Degrees_east	Tower	(Coutts *et al.*, 2007a)
3	Ground height	93	m	Tower	(Coutts *et al.*, 2007a)
4	Measurement height above ground	40	m	Tower	(Coutts *et al.*, 2007b)
5	Impervious area fraction	0.62	1	500 m radius	(Grimmond *et al.*, 2011)
6	Tree area fraction	0.225	1	500 m radius	(Grimmond *et al.*, 2011)
7	Grass area fraction	0.15	1	500 m radius	(Grimmond *et al.*, 2011)
8	Bare soil area fraction	0.005	1	500 m radius	(Grimmond *et al.*, 2011)
9	Water area fraction	0	1	500 m radius	(Grimmond *et al.*, 2011)
Detailed experiment parameters (1–24)					
10	Roof area fraction	0.445	1	500 m radius	(Grimmond *et al.*, 2011)
11	Road area fraction	0.13	1	500 m radius	(Grimmond *et al.*, 2011)
12	Other paved area fraction	0.045	1	500 m radius	(Grimmond *et al.*, 2011)
13	Building mean height	6.4	m	500 m radius	(Grimmond *et al.*, 2011)
14	Tree mean height	5.7	m	500 m radius	(Nice *et al.*, 2018)
15	Roughness length momentum	0.4	m	500 m radius	(Coutts *et al.*, 2007b)
16	Displacement height	7.92	m	500 m radius	(Coutts, 2006, p. 228)
17	Canyon height width ratio	0.42	1	500 m radius	(Grimmond *et al.*, 2011)
18	Wall to plan area ratio	0.4	1	500 m radius	(Grimmond *et al.*, 2011)
19	Average albedo at midday	0.151	1	Radiometer view	Median of observations
20	Resident population density	2,940	Person km^{-2}	Suburb average	(Coutts *et al.*, 2007a)
21	Anthropogenic heat flux mean	11	$W{\cdot}m^{-2}$	500 m radius	(Best and Grimmond, 2016a)
22	Topsoil clay fraction	0.18	1	250 m grid	(Hengl, 2018a)
23	Topsoil sand fraction	0.72	1	250 m grid	(Hengl, 2018b)
24	Topsoil bulk density	1,230	Kg m^{-3}	250 m grid	(Hengl, 2018c)

Note: Parameters 1–9 were provided to participants for use in the 'baseline' experiment, while the 'detailed' experiment allowed the use of all parameters. For detailed definitions see Lipson *et al.* (2022a).

BEPCOL, K-UCMv1, TARGET) because a long spin-up was not deemed necessary by those participants.

2.2.4 | Baseline and detailed experiments

To assess how site-specific information impacts model performance, two experiments are undertaken. First, as a *baseline*, participants configured their models using only land cover and location information (parameters 1–9, Table 3) and their default model configurations. This is designed to evaluate models configured with information typically obtainable from global high-resolution land cover datasets. Second, for a *detailed* submission, participants could use all parameters in Table 3. This is designed to evaluate if performance improves with parameters that are more challenging to obtain and not typically globally available (e.g., building height, canyon aspect ratio, and a breakdown of hard surfaces into building, road and paved fractions).

The previous intercomparison at the same site (PILPS-Urban: Grimmond *et al.*, 2011) included four stages with increasingly detailed site information for participants. The baseline experiment in the current project is most similar to PILPS-Urban Stage 2, and the detailed experiment to PILPS-Urban Stage 4 (for which model

TABLE 4 Observational data description and availability

Variable	Description	Units	Positive	All	Day	Night	DJF	MAM	JJA	SON
a. Forcing data				[%]	[%]	[%]	[%]	[%]	[%]	[%]
SW_{down}	Downward short-wave radiation	$W \cdot m^{-2}$	Downward	85.7	38.8	47.7	19.2	19.4	19.1	28.0
LW_{down}	Downward long-wave radiation	$W \cdot m^{-2}$	Downward	71.8	38.2	33.6	19.2	19.4	13.7	19.5
T_{air}	Air temperature	K	-	100.0	52.3	47.7	19.2	19.4	23.5	37.9
Q_{air}	Specific humidity	$kg \cdot kg^{-1}$	-	100.0	52.3	47.7	19.2	19.4	23.5	37.9
P_{surf}	Station air pressure	Pa	-	86.3	46.0	40.2	19.1	19.1	16.4	31.7
Wind_N	Northward wind component	$m \cdot s^{-1}$	Northward	99.9	52.2	47.7	19.2	19.4	23.5	37.9
Wind_E	Eastward wind component	$m \cdot s^{-1}$	Eastward	98.9	51.7	47.3	18.8	19.4	23.4	37.2
Rainf	Rainfall rate	$kg \cdot m^{-2} \cdot s^{-1}$	Downward	100.0	52.3	47.7	19.2	19.4	23.5	37.9
Snowf	Snowfall rate	$kg \cdot m^{-2} \cdot s^{-1}$	Downward	0.0	0.0	0.0	0.0	0.0	0.0	0.0
b. Analysis data										
SW_{up}	Upward short-wave radiation	$W \cdot m^{-2}$	Upward	35.9	35.9	0.0	11.5	7.3	5.9	11.2
LW_{up}	Upward long-wave radiation	$W \cdot m^{-2}$	Upward	66.2	35.1	31.1	19.0	15.8	13.4	18.0
Q_{le}	Latent heat flux	$W \cdot m^{-2}$	Upward	43.1	19.8	23.4	11.3	9.3	11.5	10.9
Q_h	Sensible heat flux	$W \cdot m^{-2}$	Upward	43.3	19.8	23.5	11.3	9.4	11.6	11.0
Q_{tau}	Momentum flux	$N \cdot m^{-2}$	Downward	73.3	35.4	37.9	18.9	15.8	14.1	24.5

Note: Forcing data are gap-filled with bias-corrected reanalysis data (Lipson *et al.*, 2022a). Analysis data are used for model evaluation without gap-filling. DJF=December, January, February (summer); MAM: March, April May (autumn); JJA: June, July, August (winter); SON: September, October, November (spring).

outputs are reanalysed and compared with the current cohort in Section 3: Results).

2.2.5 | Requested model outputs

Of the 55 variables participants are asked to return, here we analyse four surface energy fluxes – upward short-wave (SW_{up}) and long-wave (LW_{up}) radiation, sensible (Q_h) and latent heat flux (Q_{le}) – as well as the momentum flux (Q_{tau}) (Table 4b). The additional 50 variables are collected to undertake more detailed analysis in future studies and for error checking purposes (e.g., to check input forcing aligned with output time steps).

Variable names and formats follow the conventions of the Assistance for Land-surface Modelling Activities (ALMA) (Bowling and Polcher, 2001), as used in previous PILPS projects to facilitate data exchange in (non-urban) land surface model intercomparisons projects. Variables requested include both the ALMA 'mandatory' and additional urban-specific variables [e.g., anthropogenic heat (Q_{anth}) and water (Q_{irrig}) fluxes, storage heat flux (Q_{Stor}), roof, wall and road surface temperatures (RoofSurf$T$, WallSurf$T$, RoadSurf$T$), bulk air temperature within buildings and in street canyons ($T_{airBuilding}$, $T_{airCanyon}$), and urban canopy albedo (U_{Albedo})]. Outputs are further described in the modelling protocol (Lipson *et al.*, 2020). No submission included all requested outputs (Figure 4).

2.2.6 | Submission and feedback

Submissions were accepted through a web portal (https://modelevaluation.org) that stores data and undertakes comparison with observation (Abramowitz, 2018). Various automatic and manual checks (Table 5) are undertaken to diagnose human errors in model configuration and outputs, as these cause poor performance that prevents model design or parameterisation from being appropriately assessed (Menard *et al.*, 2021). On submission, immediate feedback is provided to participants to inform of basic file formatting errors. Subsequently, time series and energy closure plots are provided to participants by project coordinators. Short-wave radiation is chosen as a focus for feedback because it is a relatively simple flux to model, and has an instantaneous response, making timing issues between forcing and outputs more obvious. Also, correctly representing the bulk albedo is known to be important for urban model performance, as the net short-wave radiation dominates the energy balance (Best and Grimmond, 2015).

Following feedback, participants had an opportunity to resubmit prior to more complete analysis and final

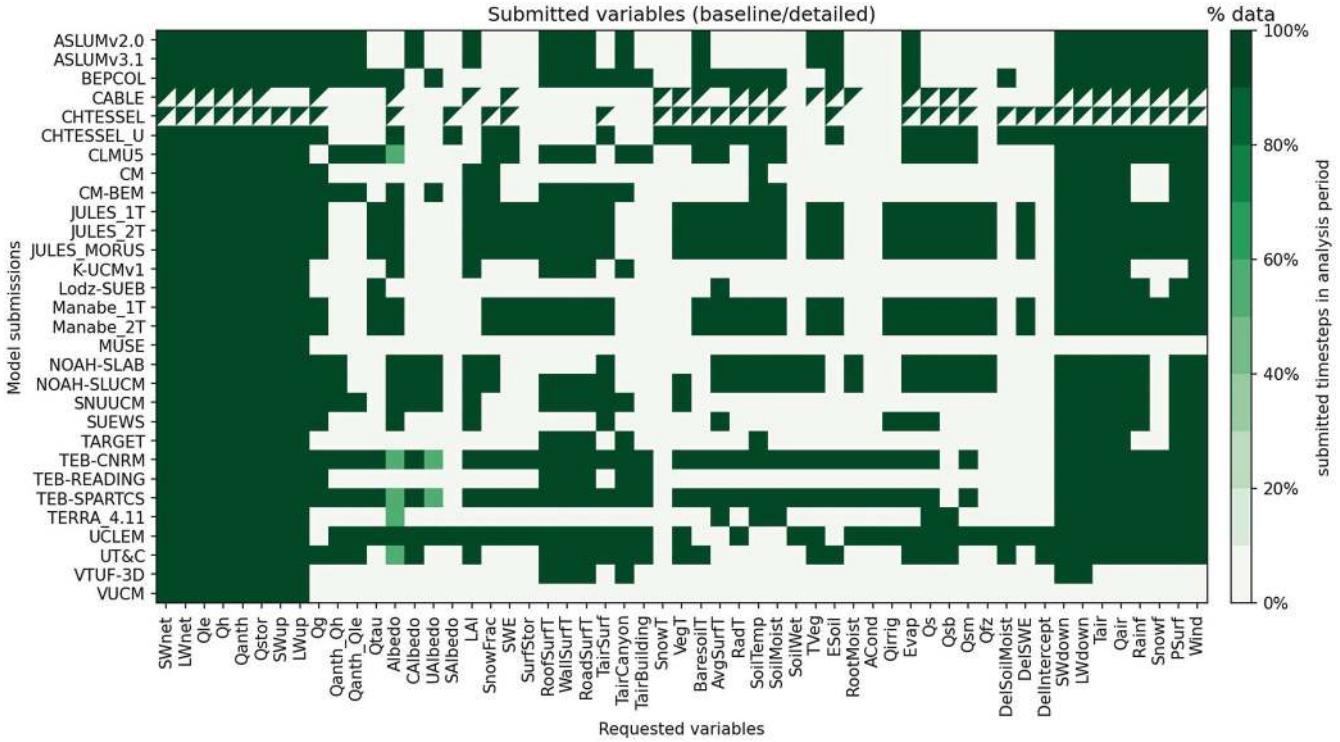

FIGURE 4 Submitted variables. Variables analysed here are defined in Table 4, with others defined in the Urban-PLUMBER modelling protocol (Lipson *et al.*, 2020), following the ALMA naming conventions (Bowling and Polcher, 2001). Two models (CABLE and CHTESSEL) submitted only the baseline experiment, as they could not use the detailed urban parameters (Table 3).

TABLE 5 Submission error checks. Feedback was provided to participants, who were able to resubmit prior to final analysis

Check action	Purpose	Number affected
a) Immediate feedback from modelevaluation.org		
Timestep number	Ensure time step length and simulation period matched expectations	Some
Included variables	Check number of variables submitted	All
Variable names	Check submitted variables names are as requested	Some
Variable units	Check submitted variable units are as requested	Some
Mean fluxes plots	Provide feedback on sign and magnitude of mean fluxes	Some
b) Feedback after manual checks (i.e., subsequent weeks): Plots include:		
SW_{down}	Ensure timestamps of submission matched expectations	Some
SW_{down} (subset)	Check if modelled SW_{down} matches forcing: Some using <30-min time steps introduces	
Interpolation errors	Some	
Energy closure	Lack of surface energy balance closure may indicate incorrect partitioning, output format or sign errors	Many
SWnet (average)	Simulated midday albedo: Midday albedo provided for detailed experiment	Some
Anthropogenic flux	Simulated anthropogenic flux: Expect mean magnitude provided for detailed experiment	Some

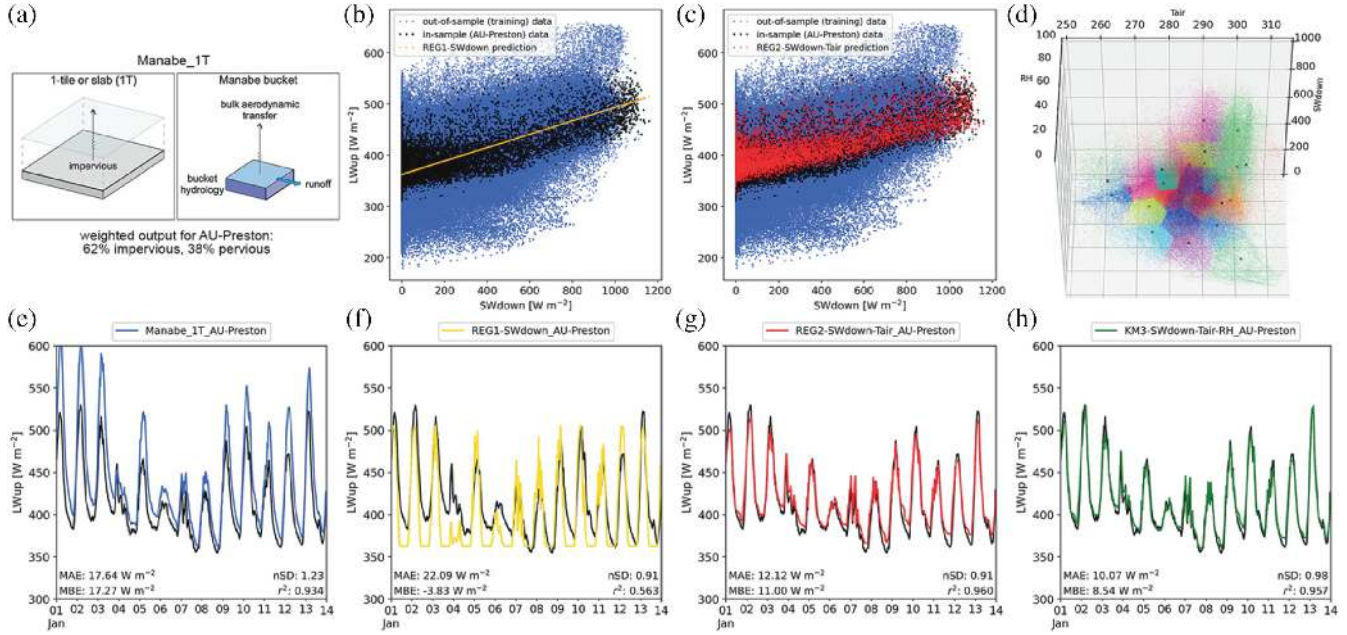

FIGURE 5 Example benchmark diagrams. (a) Manabe_1T, (b) REG1-SW_{down}, (c) REG2-SW_{down}-T_{air}, (d) KM3-SW_{down}-T_{air}-RH with cluster centroids (black crosses), and (e–h) predictions using corresponding (a–d) benchmarks. A two-week period of upward long-wave radiation (LW_{up}) is used as an example of benchmark output (coloured lines, with observations in black). Error statistics are calculated over all available periods, for mean absolute error (MAE), mean bias error (MBE), normalised standard deviation (nSD) and the coefficient of determination (r^2), all defined in Table A1.

error statistics being shared. The number of submissions (including the first) varied from one to nine. Initial checks identified human errors including incorrect start times, mislabelling of outputs, variable sign errors, forcing interpolation errors (where model time step was shorter than forcing) and errors in model source code causing unphysical or unexpected behaviour. Results were presented to participants through the project website (https://urban-plumber.github.io/), including time series plots of all submitted variables and individual model results and metadata (archived in the supporting information).

2.3 | Evaluation methods

2.3.1 | Benchmarks

Following PLUMBER (Best *et al.*, 2015), we use benchmarks (Figure 5) to guide performance expectations that are both physically (e.g., Manabe, 1969; bucket model) and empirically based. The empirical benchmarks are determined by statistical regressions using observational data independent of the site (so-called 'out-of-sample'), meaning that data from the site being tested are not used to establish regression parameters. PLUMBER used out-of-sample benchmarks to provide a lower bound on performance expectations. Here we include two 'in-sample' empirical benchmarks (i.e., derived using test site data) to give an expected upper bound on flux predictability. More complex empirical models with lagged inputs can improve benchmarks further (Haughton *et al.*, 2017); however, the benchmarks described below are sufficient for our analysis and allow direct comparison with those used in PLUMBER.

The out-of-sample regressions are trained using meteorological data: downward short-wave radiation (SW_{down}), air temperature (T_{air}), and relative humidity (RH) from 20 urban sites (Table A2). The sites are from Europe, the Americas, Asia and Australia, and have different regional climates, urban surface characteristics and observational period (Lipson *et al.*, 2022a). All empirical benchmarks rely only on contemporaneous meteorological data (i.e., do not draw on data from previous periods to make predictions).

Six benchmarks are categorised into three groups. The ALMA short-names (Table 4) are used to denote model driving variables of empirical benchmarks.

Group one includes a single *physically-based* benchmark:

- *Manabe_1T*: A simple 'slab and bucket' model (Figure 5a) based on physical principles (i.e., conservation of energy, mass and momentum). The impervious (built) fraction is simulated using a one-tile slab scheme

(Best, 2005). For the pervious fraction a simple representation allows precipitation to fill a store which overflows when full, and otherwise freely evaporates (Manabe, 1969). At each time step, the impervious and pervious tile outputs are calculated and aggregated with a weighted mean. Manabe_1T is configured using baseline parameters (Table 3: parameters 1–9). Additionally, Manabe_1T is treated as a participating model (i.e., evaluated against other benchmarks) through a secondary configuration using the detailed site parameters (Table 3).

Group two has three *out-of-sample empirical* benchmarks:

- *REG1-SW$_{down}$*: Linear regression with one variable (SW_{down}, Figure 5b) is used separately to predict SW_{up}, LW_{up}, Q_h, Q_{le}, and Q_{tau}. At night all predicted values are constant because $SW_{down} = 0\,W{\cdot}m^{-2}$.
- *REG2-SW$_{down}$-T$_{air}$*: Two-variable (SW_{down} and T_{air}) linear regression (Figure 5c) provides some information at night and provides benefit for variables strongly dependent on temperature (e.g., LW_{up} and Q_h).
- *KM3-SW$_{down}$-T$_{air}$-RH*: A piecewise multivariable regression. Following PLUMBER's conceptual arguments, three predictor variables (SW_{down}, T_{air} and RH data) are split into three groups (low, medium and high) to create $3^3 = 27$ clusters, for which independent regressions are trained. *k*-means clustering (Pedregosa *et al.*, 2011) is used to partition training data unsupervised (Figure 5d). To use this benchmark, at each time step the proximity of the forcing data to one of the 27 training cluster centroids is determined, and then that cluster's regression is applied to form a prediction. This benchmark equates to PLUMBER's EMP3KM27 (Best *et al.*, 2015), which outperformed all participating land surface models when predicting sensible and latent heat fluxes across 20 sites based on common metrics.

Group three has two *in-sample empirical* benchmarks:

- *KM3-IS-SW$_{down}$-T$_{air}$-RH*: Following the previous *k*-means clustering method, but trained with in-sample data only (i.e., AU-Preston). This will outperform an equivalent out-of-sample model, but performance is expected to degrade if applied to dissimilar conditions (i.e., another site) because of overfitting.
- *KM4-IS-SW$_{down}$-T$_{air}$-RH-Wind*: *k*-means is applied incorporating a fourth variable (wind speed), increasing the clusters to 3^4 following the above rationale. Wind speed provides information to help predict turbulent heat and momentum fluxes.

Benchmark time series data are openly available (Lipson and Best, 2022).

2.3.2 | Error metrics

Following PLUMBER (Best *et al.*, 2015), we use statistical measures in three groups:

- *Commonly used* model comparison statistics: mean absolute error (MAE) measures average error; mean bias error (MBE) for overall bias; normalised standard deviation (nSD) compares the variance of model output to that of the observations; correlation coefficient (r) measures pattern errors.
- *Extremes* of the observed distribution: absolute error at the 5th and 95th percentile of observed and modelled outputs.
- *Shape of the distribution* compared to observations: *skewness* measures differences in symmetry of the distributions; *kurtosis* measures differences in the weight of the tails of the distribution; *overlap* indicates the closeness of fit across the two distributions.

We also separately use centred root-mean-square error (cRMSE) as a measure which combines variance and pattern errors, but does not capture bias errors (Taylor, 2001). For aggregated scoring, error statistics (e.g., MBE) are redefined into error metrics (m_{MBE}) to be positive with perfect score of zero (Table A1). For benchmark scoring, MAE gives identical results to the normalised mean error (NME) used in PLUMBER. All statistics and metrics are defined in Appendix Table A1.

2.3.3 | Benchmark scoring

PLUMBER used a simple rank-based score to evaluate models and benchmarks (Best *et al.*, 2015). However, simple ranking may give a false impression of difference where metric results are nearly identical (Haughton *et al.*, 2016). Relative scoring allows relative performance to be shown (Sabot *et al.*, 2020). Furthermore, if global extrema are used across all models and benchmarks, this ensures a single benchmark has the same relative score across models.

Thus, our scoring differs from PLUMBER. For each participating model i, variable v and metric m, a score S is calculated for the i th model using the minimum and maximum metric result across all models and benchmarks for that variable:

$$S_{i,v,m} = \frac{m_{i,v} - \min(m_v)}{\max(m_v) - \min(m_v)} \quad (1)$$

This gives a score of 0 for the best-performing model or benchmark, and 1 for the poorest, with all others scaled relative to the range of results. Rescaling scores between 0 and 1 ensures that no metric has greater weight when aggregated with others. Different metric scores (Section 2.3.1) are aggregated into groups with:

$$\overline{S}_{i,v} = \frac{1}{n_m} \sum_{m=1}^{n_m} S_{i,v,m} \quad (2)$$

where n_m is the number of metrics in group being aggregated (e.g., for the extremes group $n_m = 2$).

3 | RESULTS

To build our understanding of the model's performance, we initially consider one error statistic, the MAE (Section 3.1: Figure 6). Although no single measure can fully characterise model skill (Jackson *et al.*, 2019), MAE provides a simple and unambiguous measure of average error (Willmott and Matsuura, 2005) and allows comparison of error magnitude across fluxes in natural units ($W \cdot m^{-2}$). Subsequently, three statistics (for correlation, variance and difference errors) are analysed in a Taylor diagram (Taylor, 2001) (Section 3.2: Figure 7). Aggregated benchmark performance scores (Section 2.3.3) are then analysed in benchmarking diagrams, using common error metrics (Section 3.3: Figure 8). Finally, all metrics (common, extreme and distribution) are used to compare models with benchmarks (Section 3.4: Figure 9). The PILPS-Urban Phase 2 Stage 4 (Grimmond *et al.*, 2011) (hereafter G11) anonymised model outputs are reanalysed here conforming to this project's metrics and analysed periods.

3.1 | Assessment using the mean absolute error

Individual model MAE results are combined into boxplots (Figure 6) for three experiments (G11, baseline and detailed) with results also analysed by model cohort (Section 2.1). The performance of the ensemble mean (i.e., the mean of participating model outputs at each time step; rightmost column) and the benchmarks (coloured horizontal lines) are also shown (Figure 6).

For upward short-wave radiation (SW_{up}), the detailed site information (e.g., albedo) improved all cohort performance where utilised (non-urban models did not submit detailed simulations). Most one-tile, two-tile and canyon models outperform the physical and out-of-sample benchmarks when given detailed information, whereas complex models did not use this information as effectively. The ensemble means perform similarly across the three experiments (G11, baseline and detailed), matching the best-performing individual models. The relatively low MAE for all benchmarks (2.3–7.0 $W \cdot m^{-2}$) indicates this flux can be well simulated with few inputs.

For upward long-wave radiation (LW_{up}), providing more detailed site information reduced the MAE for one-tile models, but performance changed little for other model types (in some cases becoming poorer). Most models outperform the physical benchmark but only some beat the empirical benchmarks. The ensemble mean time series outperforms the physical and out-of-sample benchmarks. The LW_{up} benchmark MAE values are larger (5.2 — 22.4 $W \cdot m^{-2}$) than for SW_{up}, indicating the flux is more challenging to predict with the information available to benchmarks.

For sensible heat flux (Q_h), providing detailed information broadly improves performance, particularly for two-tile, canyon and complex cohorts. Models with initially large baseline anthropogenic heat fluxes benefitted from knowing this site's relatively small flux magnitude (11 $W \cdot m^{-2}$ annual mean). All cohort mean and median MAE outperform the physical and out-of-sample empirical benchmarks for the detailed simulations, with the ensemble mean able to outperform all benchmarks (including in-sample benchmarks). The larger benchmark errors (18.5–32.9 $W \cdot m^{-2}$) indicates that the flux is more challenging to predict than either radiative flux.

For latent heat flux (Q_{le}), more detailed information provided little benefit for reducing MAE, and performance degrades slightly with more complex cohorts (based on built geometry, not vegetation or hydrology attributes). This suggests the more detailed information (mostly related to urban morphology) may not be in a form useful for models. However, the non-urban models do most poorly as, without any impervious surface fraction, they vastly overestimate evapotranspiration. The detailed ensemble mean outperforms all individual models and the in-sample empirical benchmarks. Q_{le} has a relatively high benchmark range (18.6–26.1 $W \cdot m^{-2}$), indicating a greater challenge to predict than radiative fluxes.

In summary, the range of MAE is lower for the current models than for the G11 models, indicating better performance of urban models in the present intercomparison. Differences in mean MAE for models in G11 and detailed experiments (which have comparable site information) are statistically significant for SW_{up}, Q_h and Q_{le}, but not for LW_{up} (t-test, $p < 0.05$). The range of MAE for the detailed simulations is generally smaller than for the baseline, indicating models benefitted from additional site information. The difference in the mean MAE for baseline and detailed experiments reaches significance in SW_{up} only.

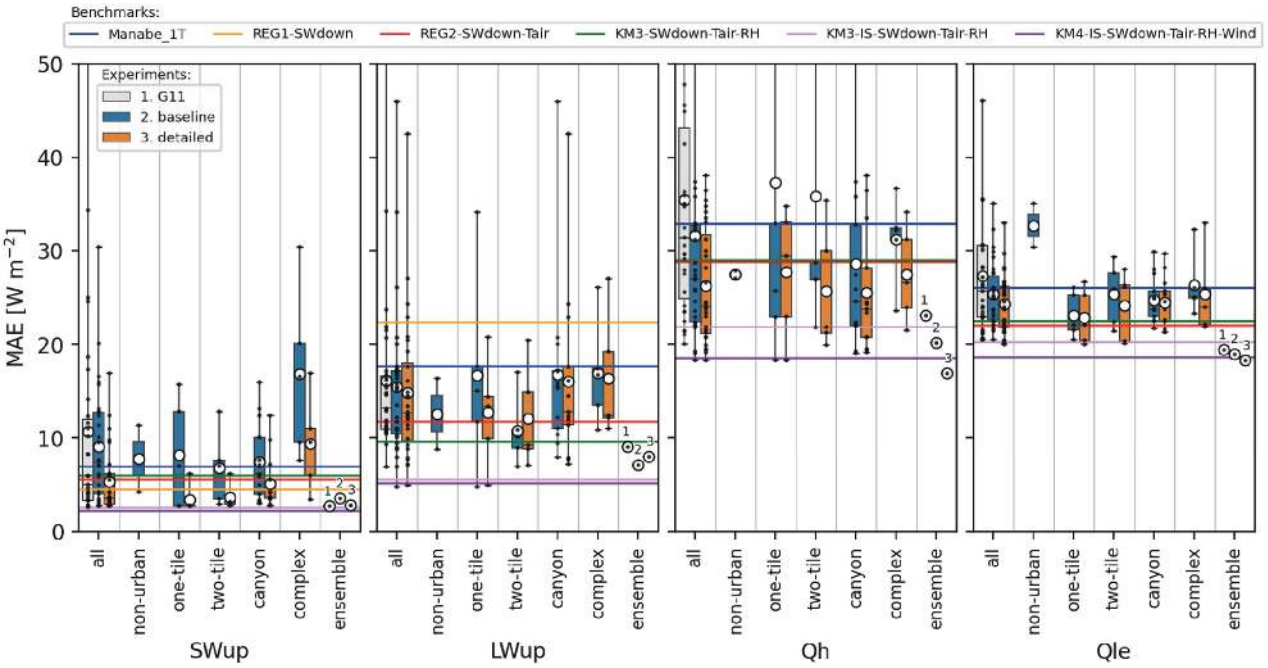

FIGURE 6 Mean absolute error (MAE) boxplot results. Individual models (dots) are split into cohorts (Section 2.1) with benchmarks (horizontal lines) for four fluxes: upward short-wave radiation (SW_{up}), upward long-wave radiation (LW_{up}), sensible heat flux (Q_h) and latent heat flux (Q_{le}). Boxplots are shown for model experiments (left to right in each column): (1) PILPS-Urban Phase 2 Stage 4 (G11: Grimmond *et al.*, 2011; grey); (2) Urban-PLUMBER baseline (blue); and (3) Urban-PLUMBER detailed (orange). The level of information in G11 (the final stage of PILPS-Urban) is comparable to the detailed experiment in this project. Boxes show the 25th and 75th percentile, median (horizontal line), full range (whiskers) and mean (open circle).

3.2 | Taylor diagram evaluation

A Taylor diagram (Taylor, 2001) combines three error statistics: (a) a difference error metric (centred root-mean-square error: cRMSE); (b) a variance error metric (modelled standard deviation normalised by observed standard deviation: $\tilde{\sigma}$); and (c) a correlation error metric (Pearson's correlation coefficient: r; all defined in Table A1). Taylor diagrams use the centred RMSE (the RMSE after mean bias is removed) because it has a geometric dependence with the other two (independent) metrics, allowing the construction of the diagram (Figure 7). For each model, a marker shows where the three metrics intersect. The cRMSE of benchmarks is indicated by the concentric dashed lines. A model that would perfectly align with observations is indicated with a star at the figure base. The G11 PILPS-Urban Phase 2 Stage 4 results (small dots) are compared with the detailed experiments, as the site information available in each is comparable.

For SW_{up}, most Urban-PLUMBER (UP) models and benchmarks are grouped tightly around the observation star (Figure 7a). Some UP models (e.g., 22, 29) have high correlations, but different variances than observed, indicating errors in bulk albedo. Others (e.g., 09) captured the observed variance well, but had lower correlation, indicating a potential time-of-day issue with SW_{up} (either a time offset, or in this model's case, an asymmetrical diurnal profile). Using cRMSE as a metric, 23 of 30 UP models outperform at least one benchmark, while only 18 of 31 of the G11 models do (Table 6). The spread in $\tilde{\sigma}$ and r indicates G11 models had greater albedo and time-of-day errors than UP models (Figure 7).

For LW_{up} (Figure 7b), many participating models have larger variance than observed because of an overprediction in the diurnal range of LW_{up}. Some UP models (e.g., 22, 20) are high-end outliers, with $\tilde{\sigma}$ of approx. 1.7 (i.e., 170% of observations), greater than the cRMSE of all benchmarks. G11 outliers are larger still (0.5 to 2.0). The LW_{up} ensemble mean in UP and G11 performs similarly, as do the number of models that outperform benchmarks (Table 6). In UP, one model (18: NOAH-SLAB) outperformed all benchmarks in cRMSE.

For Q_h (Figure 7c), correlation and variance statistics (and hence cRMSE) improved substantially in UP compared with G11. For UP, 28 of 30 models outperform at least one benchmark in cRMSE (cf. 18 of 31 in G11, Table 6). Twelve UP models outperform all out-of-sample benchmarks (cf. six G11 models). One UP model (14: Lodz-SUEB) outperforms the four-variable in-sample benchmark, which is the upper limit of performance expectations. The UP ensemble mean also performs very well, outperforming all benchmarks nSD, R and cRMSE.

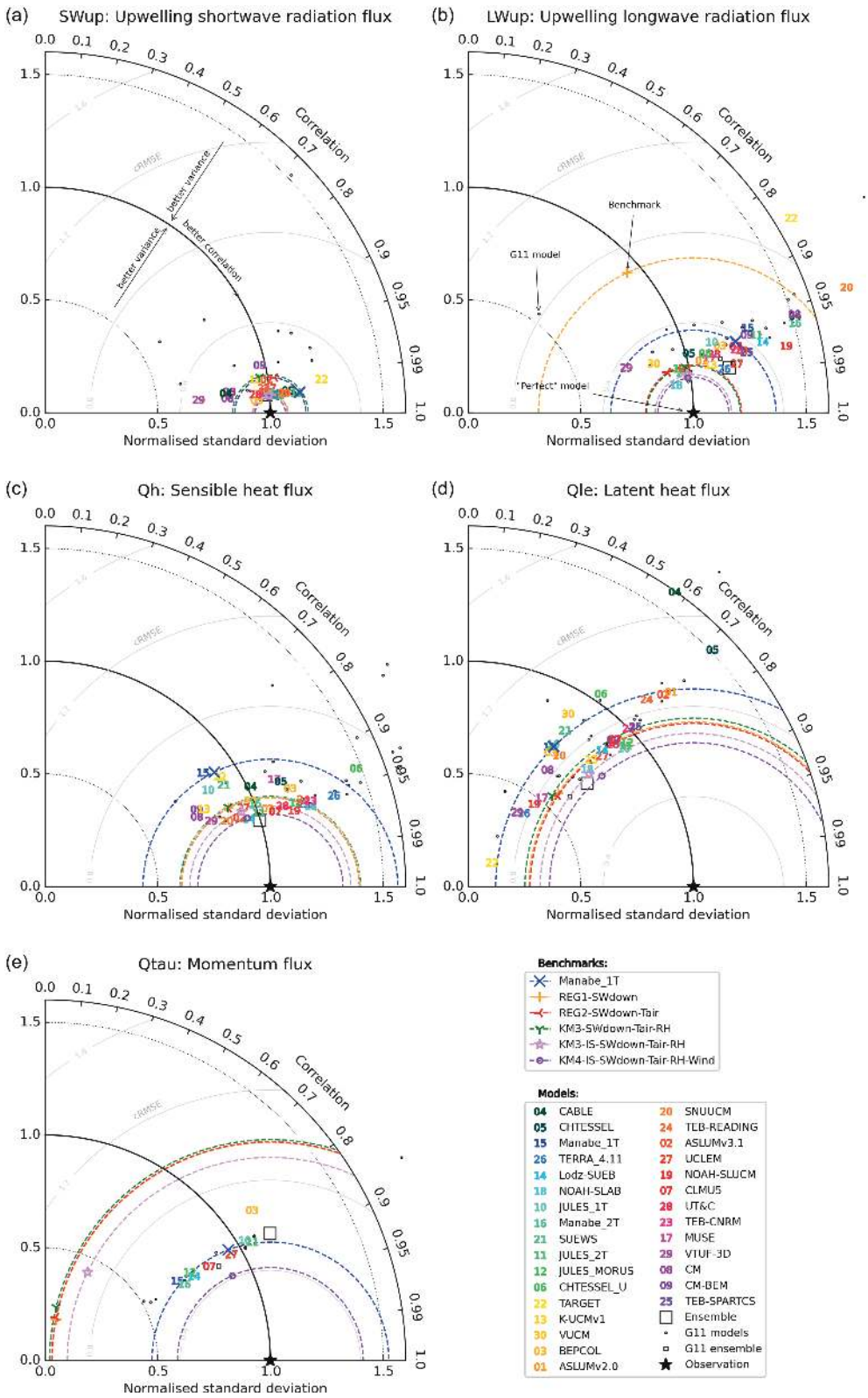

FIGURE 7 Taylor diagram combines the normalised standard deviation ($\tilde{\sigma}$), correlation coefficient (r) and cRMSE (defined in Table A1). Models (coloured numbers) have better performance if closer to star at diagram base, with cohort colours: non-urban (dark greens), one-tile (blues), two-tile (greens), canyon (orange to reds), complex (purples). Benchmarks models (coloured symbols) with their cRMSE contours (concentric dashed lines), and the PILPS-Urban Phase 2 Stage 4 (G11: small black circles).

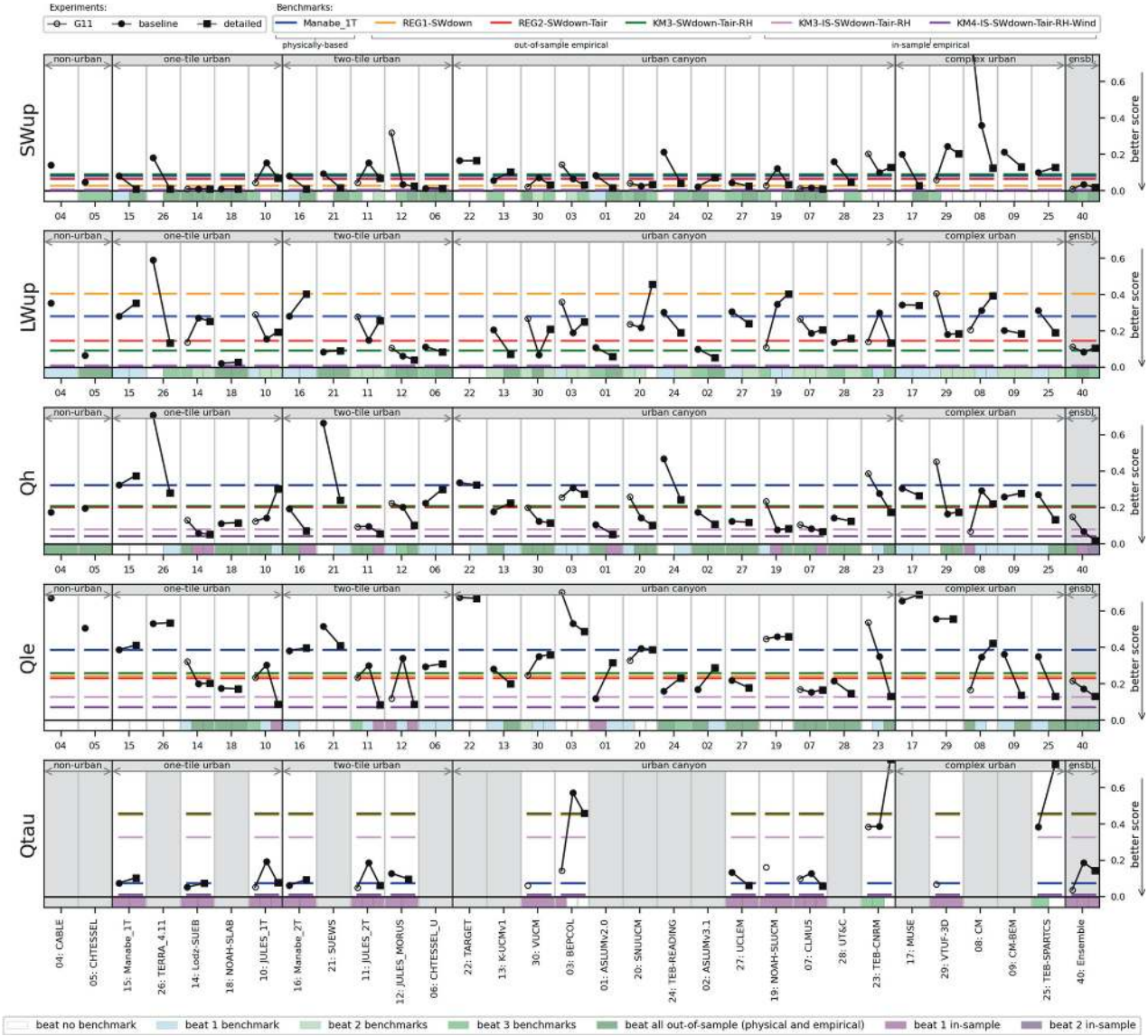

FIGURE 8 Benchmarking assessment for 'common' set of metrics (MAE, MBE, nSD, R) showing benchmarks (coloured lines) and models (black markers). Lower scores are better (Equation 1 and Equation 2). Model columns include up to three markers for different experiment submissions (1, PILPS-Urban Phase 2 Stage 4 [G11]; 2, baseline; and 3, detailed). Colours at the base of each column indicate the benchmarks a model outperforms per experiment (colour, lower legend). Models are ordered by nominally increasing complexity (Figure 2). For Q_{tau}, grey indicates no submission.

The UP ensemble mean variance is nearly identical to observations ($\tilde{\sigma} = 1.00$), while for G11, 26 models had higher variance than observations and the ensemble mean $\tilde{\sigma} = 1.23$. Higher variance in G11 models indicates an overprediction in maximum Q_h values or a general overestimation of the variability in Q_h. As cRMSE is 'centred' it does not measure bias error (Taylor, 2001). For Urban-PLUMBER, the MBE for Q_h ensemble mean is 5.2 W·m^{-2} (cf. 12.1 W·m^{-2} in G11), indicating the UP models have improved partitioning of available energy into Q_h.

Compared with other fluxes, the poorer cRMSE of the six benchmarks for Q_{le} (Figure 7d) indicates this flux is more challenging to predict, or that it requires other inputs to improve performance (e.g., precipitation, soil states or vegetation characteristics). Most UP and G11 models underestimate the variance of this flux. The ensemble mean's $\tilde{\sigma} = 0.70$ (i.e., 70% of observations standard deviation). However, this is an improvement over the G11 ensemble mean ($\tilde{\sigma} = 0.60$). The G11 results exclude six models that did not provide Q_{le} output (i.e., some assumed $Q_{le} = 0$ W·m^{-2}). The ensemble mean for Urban-PLUMBER MBE (−4.1 W·m^{-2}) is better than for G11 (−8.2 W·m^{-2} for G11 models that explicitly resolved Q_{le}). Combined with the improved ensemble mean Q_h MBE, this indicates the UP models are better at partitioning available energy

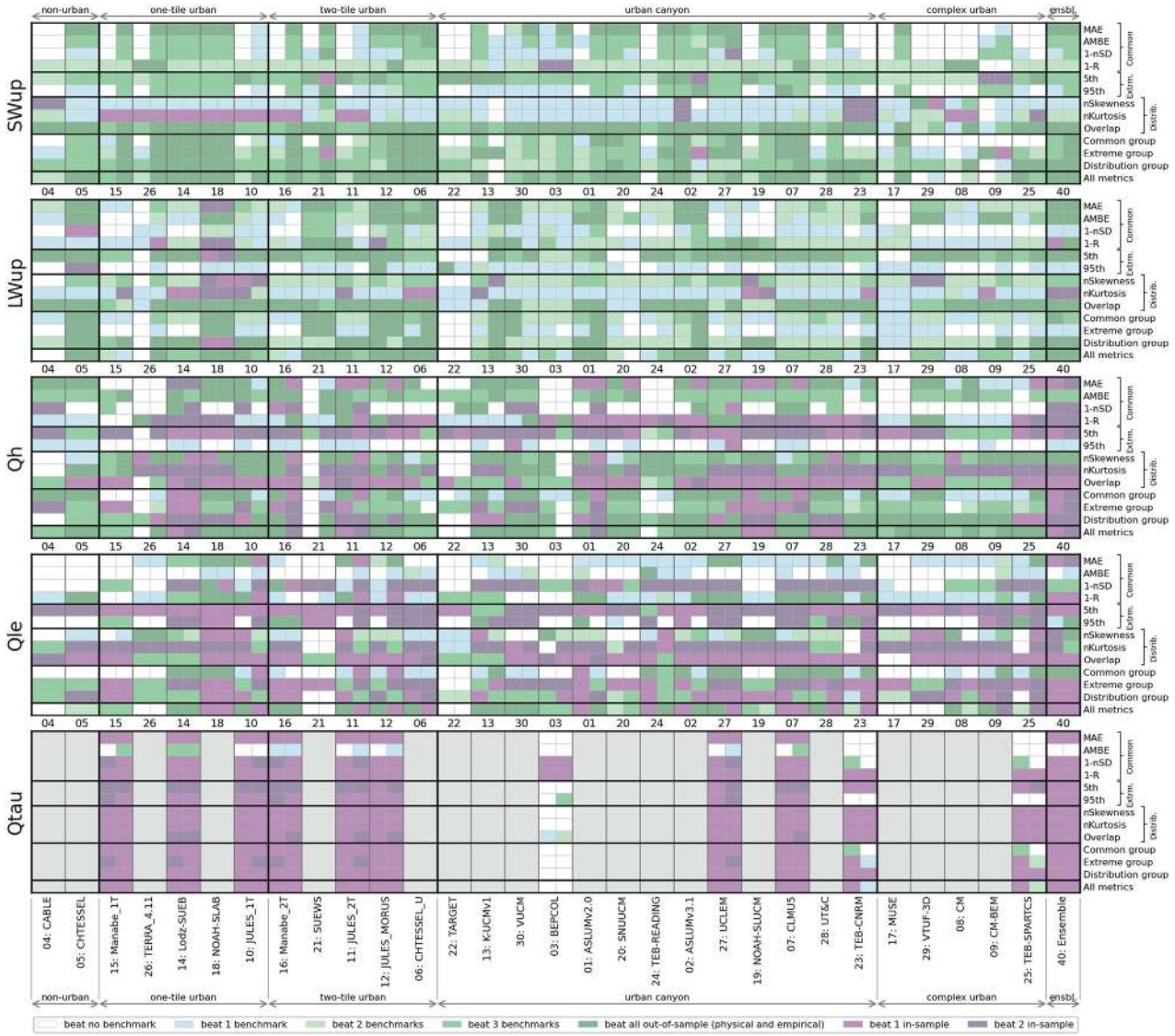

FIGURE 9 All metric benchmarking results. Models are ordered nominally by increasing complexity (Figure 2) showing benchmark scorecard of individual and aggregated metrics based on ability to outperform benchmarks (colour) for both the baseline (left) and detailed (right) simulations. Note a single box indicates only one submission made. For Q_{tau}, grey indicates no submission.

TABLE 6 Number of models outperforming the cRMSE benchmarks for Urban-PLUMBER detailed simulations (UP) and PILPS-Urban Phase 2 Stage 4 (G11) (Grimmond *et al.*, 2011)

	Flux	SW_{up}		LW_{up}		Q_h		Q_{le}		Q_{tau}	
	Project	UP	G11	UP	G11	UP	G11	UP	G11	UP	G11
	Total models	30	31	30	31	30	31	30	25	11	11
Beat at least:	1 benchmark (physical or empirical)	**23**	18	28	28	**28**	18	**22**	16	9	**10**
	2 benchmarks (physical or empirical)	**22**	18	**20**	16	**15**	6	**10**	7	9	**10**
	3 benchmarks (physical or empirical)	**22**	18	4	**6**	**12**	6	**8**	5	9	**10**
	All out-of-sample benchmarks (physical and empirical)	0	**1**	4	**6**	**12**	6	**5**	3	**4**	3
	1 in-sample empirical benchmark	0	0	1	1	**7**	2	0	0	9	**10**
	2 in-sample empirical benchmarks	0	0	1	1	**1**	0	0	0	0	0

Note: Models are executed for the same site with the same observations but different models (or versions). Higher number (bold) is better.

into Q_h and Q_{le}. No model in either project outperformed in-sample empirical benchmarks for Q_{le} (Table 6).

Eleven models provided the simulated momentum flux (Q_{tau}) from both Urban-PLUMBER and G11 (Figure 7e). Performance in both projects is similar, cf. benchmarks (Table 6). Benchmarks without wind information (i.e., one-, two- and three-variable empirical benchmarks) did not perform well. All models ranked between the three-variable and four-variable in-sample benchmarks, and all were able to beat out-of-sample empirical benchmarks.

3.3 | Benchmarking evaluation: Common metrics

We can evaluate model results relative to the benchmarks for various experiments using the aggregated scores (Equation 1 and 2) from four common metrics (MAE, MBE, nSD, R). A lower relative score indicates better performance (Figure 8). Of the 30 models in this project, 11 (in an earlier form) also participated in G11, so allow direct comparison. In Figure 8, the models are ordered in increasingly complex cohorts (Section 2.1), and within cohorts by the 'total complexity' (Figure 2), which includes hydrology and anthropogenic related characteristics.

For SW_{up}, providing more detailed site information consistently improves the aggregate scores. The models in the simpler cohorts (one-tile, two-tile) benefit more from the more detailed information (square marker in each model column), where eight of 10 outperformed both physical and out-of-sample empirical benchmarks (dark green, model column base). Almost all canyon models outperform a benchmark, and in the detailed experiment three outperform all physically-based and out-of-sample empirical benchmarks. Only one complex model outperforms a benchmark for SW_{up} (excluding G11 results), indicating the complex cohort have more difficulty using provided site information. The ensemble mean (last column) for G11 and UP performs well, beating all out-of-sample benchmarks (dark green) in the more detailed experiments.

For LW_{up}, additional site information does not always improve performance, and sometimes degrades it. Performances of cohorts are inconsistent, with some models in non-urban, one-tile, two-tile and canyon categories outperforming all physical and out-of-sample empirical benchmarks, but others beat none. The canyon and complex models, and some two-tile models with radiation parameterisations, should be able to improve their LW_{up} performance by utilising the detailed information provided on building morphology (e.g., representing canyon long-wave trapping), but appear unable to do so. The most complex urban schemes again are not able to effectively use provided information, with none outperforming a multivariate empirical benchmark. The ensemble mean performance changes little between G11 and UP.

For Q_h, the non-urban, and most one-tile, two-tile and canyon models outperform all out-of-sample empirical benchmarks (dark green). This is in stark contrast to PLUMBER (Best *et al.*, 2015) when no model outperformed even the one-variable linear regression at non-urban sites. Again, fewer complex models are able to beat empirical benchmarks, but all outperform the physically-based benchmark. Some one-tile (14), two-tile (11, 16) and canyon (01, 07, 19) models beat the three-variable, but not the four-variable in-sample benchmarks, which we consider as an upper performance expectation. Providing additional site information more frequently improves, rather than degrades, performance. The biggest improvements occur in models that assumed initially large baseline anthropogenic heat fluxes (e.g., models 21, 24, 26), drawing on the detailed characteristic estimates of mean anthropogenic flux. Substantial improvement in the ensemble mean occurs from G11 to UP baseline to detailed experiments, with the latter outperforming all benchmarks, including those trained in-sample (purple). Q_h is the only variable where the ensemble mean beats all the benchmarks for common metrics.

Although additional site information degrades Q_{le} model performances as often as it improves it, the improvements are larger in magnitude. Hence, the ensemble mean improves across the three experiments, in each case outperforming all the out-of-sample and physical benchmarks. Non-urban models are unable to outperform any benchmark because they overestimate Q_{le} magnitudes. The three best-performing models in Q_{le} (10, 11, 12) are all JULES-based models used with different urban schemes. Some one-tile (10, 14, 18), two-tile (11, 12), canyon (07, 13, 23, 27, 28) and complex (09, 25) models outperform all physical and out-of-sample benchmarks in detailed simulations. It is unclear at this stage why these models are performing well in Q_{le}, as they all differ in their approaches in representing vegetation and hydrology processes. Further analysis across multiple sites may be informative. No detailed information for vegetation is provided that may have improved model performance further, e.g., leaf area index (LAI) phenology or stomatal conductance.

The 11 models that submitted Q_{tau} results perform better for the detailed than the baseline simulations but have a slightly degraded ensemble mean than for G11. Of the benchmarks, the four-variable in-sample benchmark (i.e., including wind speed) performs best, followed by the physically-based benchmark (Manabe_1T). Other benchmarks, without lacking the critical wind speed

information for the momentum flux, perform poorly. No clear pattern is evident by modelling approaches.

3.4 | Benchmarking evaluation: All metrics

The prior metrics (MAE, MBE, nSD, R and cRMSE) are focussed on central tendency rather than the extremes and/or the distribution. Ability to predict high-impact weather events (historical, current, future climate) makes performance skill for extremes important. We therefore expand the 'common' benchmark scorecard results from the base of Figure 8 with 'extreme' and 'distribution' error metrics in Figure 9. Relative performances (as in Figure 8) for metric groups are provided in the supporting information (Figures S1–S3).

For Q_h and Q_{le}, some models outperform in-sample benchmarks (purple), with many able to outperform all out-of-sample benchmarks (green). An exception is that all model cohorts find predicting the 95th percentile for Q_h challenging (white or blue) because they overestimate the upper Q_h tail. For Q_{tau}, models generally perform very well compared with in-sample empirical benchmarks, although this flux is heavily reliant on instantaneous wind information which is only provided in the most complex benchmark (KM4-IS-SW_{down}-T_{air}-RH-Wind). Most models assessed for Q_{tau} perform similarly to the simple physically-based benchmark (Figure 8).

The ensemble mean time series performs strikingly well across all fluxes, beating in-sample benchmarks for many metrics in Q_h, Q_{le} and Q_{tau}, and beating most out-of-sample benchmarks in SW_{up} and LW_{up}. The ability of the ensemble mean to outperform even empirical models trained in-sample suggests the participating models adequately span the range of uncertainty from the parameterisation of processes.

4 | DISCUSSION

PILPS-Urban established that correctly representing the ratio of impervious to pervious surfaces is of first-order importance for urban model performance (Grimmond *et al.*, 2011), so we do not investigate that here and provide land fraction information in the first 'baseline' experiment (Table 3a). We instead assess the impact of secondary information (e.g., three-dimensional morphology, bulk albedo, anthropogenic heat) in a 'detailed' experiment (Table 3b). We also compare outputs from Urban-PLUMBER (UP) models with those from PILPS-Urban Phase 2 Stage 4 (G11) (Grimmond *et al.*, 2011), which used the same site and observations.

Current models show reduced errors across four energy fluxes (Figure 6), with a lower MAE range, and lower mean MAE for both the baseline and detailed experiments compared with G11 (however not reaching statistical significance for LW_{up}). For cRMSE (Table 6), we find broad improvement for upward short-wave (SW_{up}), sensible (Q_h) and latent (Q_{le}) heat fluxes, but little or no improvement in upward long-wave (LW_{up}) and momentum (Q_{tau}) fluxes. When assessing performance using four common evaluation metrics (Figure 8), the ensemble mean has clearly improved for Q_h and Q_{le} but is little changed or slightly degraded for SW_{up}, LW_{up} and Q_{tau}.

These results suggest the current generation of models is performing better than the G11 models for Q_h and Q_{le}. In the last decade considerable community effort has been applied to improve existing and develop new models with particular focus on better resolving vegetation and soil processes after these were found to be highly important for performances in previous intercomparisons (Grimmond *et al.*, 2010, 2011). This implies the better performance seen here is from model development, but model application (i.e., configuration) may have also improved. Compared with G11, participants' previous experience modelling the site, the provision of more site-specific data (Table 3), the additional rapid automatic feedback identifying human errors, and/or improved spin-up strategy may also have enhanced performance, rather than model parameterisation improvements alone.

The poorer correlation of G11 models compared with current models (Figure 7) indicates time-of-day errors (i.e., human rather than model errors). However, when all models with SW_{up} correlations below 0.99 are excluded from the analysis (i.e., retaining only those with good timing), G11 models still perform poorer for LW_{up}, Q_h and Q_{le} compared with the current cohort (Figure S4), implying other factors have led to performance improvements.

Feedback provided to participants prior to final evaluation (Table 5) significantly reduces SW_{up} and LW_{up} errors for some models, and to a lesser degree also reduced Q_h and Q_{le} errors (Figure S5). Models that resubmitted more times are generally able to improve their performances relative to their first submission but did not typically outperform those that submitted only once. When models are categorised by previous experience modelling with this site (e.g., via the G11 project), more experienced groups tend to have lower errors, particularly in the initial submissions (Figure S6). This suggests that resubmission helped 'level the playing field' for those with less experience with the application of a model at this site.

Those models that undertook the extended 10-year spin-up tend to have better performance in detailed simulations than those that did not (Figure S7), however, this effect was small. Models with a closed surface energy

budget have better performance in radiative fluxes, but not in turbulent fluxes (Figure S8). Fully separating the various influences (better models, more experienced modellers, better spin-up strategy) will require additional investigation to assess their relative impacts. However, in synthesis, urban model performance has improved since the last major urban model intercomparison over a decade ago.

Flux magnitudes and dominant processes vary diurnally, and so separately analysing day and night periods provides additional insight. Compared with standard benchmarking scores (Figure 8), daytime only (Figure S9) and nighttime only (Figure S10) results are presented in the supporting information. For LW_{up} most models' daytime results degrade in comparison with benchmarks, with fewer outperforming the two-variable out-of-sample regression. However, at night, most models beat all out-of-sample benchmarks. The complex model cohort benefits most at night, with all but one complex geometry model beating all out-of-sample benchmarks, and nearly matching the performance of the in-sample benchmarks. Best and Grimmond (2015) found a similar result, concluding that the more complex geometric representations were able to account for nighttime long-wave trapping between urban structures that simpler schemes could not. In this project, some models in every cohort perform well for LW_{up} at night, but the most complex cohort was most consistently improved, and had the best overall detailed experiment results, implying they were better able to use the additional morphology information provided at that later stage. For Q_h, Q_{le}, and Q_{tau} most models across all cohorts perform very well at night, typically beating one or both in-sample benchmarks, and nearly every model easily surpassing the out-of-sample empirical benchmarks. For nighttime Q_h, some models in the urban canyon and complex categories performed better than all non-urban, one-tile and two-tile models, again implying more complex geometry is beneficial at night. The same consistent benefit from more complex geometry was not apparent in daytime periods, nor in the overall assessments.

The use of benchmarking helps to guide performance expectations (Best *et al.*, 2015). For example, without benchmarking Q_{le} appears to be poorly modelled according to the Taylor plot statistics (Figure 7), as was concluded in the earlier PILPS-Urban study (Grimmond *et al.*, 2010, 2011). However, the benchmark results show that it is more challenging for models to minimise the Q_{le} errors than, for example, SW_{up} errors (Figure 7). Benchmark assessment of extremes and distribution skill finds Q_{le} to be one of the better modelled fluxes, with models able to outperform the in-sample empirical benchmarks in many instances (Figure 9: purple). Likewise, Q_h is well modelled for the common (Figure 8) and other (Figure 9) metrics, compared to benchmarks. Fewer models outperform benchmarks for LW_{up} (Figures 7 and 8), particularly in the daytime (Figure S9). More site information generally degraded the ensemble mean skill in LW_{up} (Figures 6 and 8), except for more complex models at night (Figure S10). The overall poor performance in LW_{up} compared with benchmarks indicates an area for which model development may prove beneficial. This may be difficult, as LW_{up} is dependent on surface temperature, itself a result of the surface energy balance, so is sensitive to errors in all other surface energy fluxes (and related parameterisations). The evaluation of LW_{up} is additionally complicated by the fact that the footprints of radiative observations from a flux tower differ from the footprint of the turbulent fluxes (Schmid *et al.*, 1991; Sailor, 2011), so may be poorly represented by site parameters, which were intended to capture the larger turbulent-flux footprint.

A key PILPS-Urban (Grimmond *et al.*, 2010, 2011) finding was that simpler models generally performed as well or better than more complex models. Similarly, we find the 'complex' models (Section 2.1) are often outperformed by one-tile, two-tile and canyon models (Figures 6–9). However, model complexity needs to consider many aspects of urban environments, including morphological, hydrological, vegetation and anthropogenic influences. Some of the most complex 'built' representations have the simplest soil, water and vegetation approaches (Figure 2). The simpler models within a cohort (i.e., left side of each cohort, Figures 8 and 9) often had poorer intracohort results. Thus, the hydrological and vegetation complexity is important. Many simpler PILPS-Urban built schemes benefitted from being coupled to more sophisticated vegetation land surface models, performing well, as they do here. However, the two participating non-urban models (with sophisticated hydrology) significantly overpredict Q_{le}, performing worse than benchmarks and all other model cohorts in this flux (Figures 6–8). This indicates the representation of impervious surfaces is important even at this suburban site. Canyon models have improved compared with earlier evaluations, with some performing here as well as the best one-tile and two-tile schemes. This implies that community efforts in model development over the last decade have paid dividends, particularly the focus on integrating soil hydrology and/or vegetation into canyon models.

In stark contrast with PLUMBER (Best *et al.*, 2015), this project's submissions are often able to outperform all out-of-sample empirical benchmarks for Q_h and Q_{le}, and in some cases for the three- or four-variable in-sample benchmarks (Figures 6–9). For common metrics, PLUMBER (Best *et al.*, 2015) found no model able to outperform a single-variable linear regression using SW_{down} for Q_h, or a three-variable regression for Q_{le} in applications (over non-urban terrain). An explanation

such as Urban-PLUMBER simply having better models than PLUMBER is unlikely as some models (CABLE, CHTESSEL) or their vegetation components (NOAH, JULES) participated in both projects. Analysis coding errors are unlikely as we confirmed we could recreate the PLUMBER results using their site, model data and aggregation methods. These results suggest models for urban areas perform better than those for non-urban areas when assessed against the same empirical benchmarks.

Urban sites are highly diverse, and only one case is considered here. The AU-Preston site could be unrepresentative of other urban sites used for training, leading to poorer performing regression using out-of-sample data. This is supported by the fact that the out-of-sample three-variable regression (KM3-SW_{down}-T_{air}-RH) performed poorer than simpler regressions for Q_h and Q_{le} (Figures 6–8), indicating overfitting. However, some models outperform the regressions trained in-sample (i.e., using only AU-Preston data), and therefore good model performances are not simply related to the site's (un)representativeness. Alternatively, models may have performed well here because we provide participants with more site-specific information (Table 3) than in PLUMBER. For the latter, participants were provided with a single plant functional type descriptor (e.g., grassland). However, some Urban-PLUMBER models are outperforming benchmarks in the 'baseline' experiment when only minimal surface information is given, so better model performance is not simply from the surface descriptions provided in this project.

Ultimately, models participating in Urban-PLUMBER are performing better against benchmarks than the PLUMBER project land surface models were able to. This implies the complexity of urban surfaces benefits from the more complex modelling techniques used to address urban areas, compared with the natural landscapes evaluated in PLUMBER. A multisite evaluation is required to confirm these initial results (now under way).

In Urban-PLUMBER, we focus on bulk local-scale surface–atmosphere exchanges as these variables and scale act as the lower boundary conditions for weather, climate and air quality modelling. They may also act as the upper boundary conditions for more detailed models used for applications in cities (e.g., pedestrian thermal comfort). Some modellers using the latter type of models declined to participate in this project as their models require more detailed surface information than we could provide, and are more computationally intensive, making long (e.g., 10+ years) simulations unfeasible. Some models are not intended to predict the bulk land–atmosphere exchanges assessed here, but for predicting other details within the urban canopy. Other MIPs have encountered this challenge (bulk vs detail). ESM-SnowMIP (Menard *et al.*, 2021) found comparatively complex models developed for specific purposes, and tested rigorously for their intended use, are outperformed by simpler bulk models when bulk variables are assessed. Thus, intended model use is a key consideration when evaluating performance.

Following ESM-SnowMIP (Menard *et al.*, 2021) and our earlier experience (e.g., Grimmond *et al.*, 2010), that human errors can be widespread in intercomparison projects, we provide rapid automatic checks with feedback to participants, and follow up with manual checks (Table 5). Allowing resubmission where human errors are identified enables this evaluation to focus more closely on intended model performance. Identified human errors included: start times, output labels, variable sign, and forcing interpolation errors. These, plus model source coding mistakes, all impacted initial results. Our initial feedback focussed on SW_{up} to check that forcing and output timing aligned, and we link this to the net improvement in SW_{up} performance seen (cf. G11, Figure 6). Best and Grimmond (2015) previously established that correctly modelling the bulk surface albedo is critical for model performance for all surface energy fluxes. Ensuring albedo is simulated better in this project helps focus evaluation on other aspects of model design, such as the impact of hydrology, vegetation and anthropogenic influences.

While considerable efforts are undertaken to compare models rather than users, the different application of models will impact results, and undoubtedly some human errors remain. Hence, individual model results presented here should be interpreted with caution. We highlight broad patterns, but cannot untangle whether individual model performances are a result of: (a) aspects of model design; (b) user model configuration, or 3) model assessment methods (e.g., variables, metrics, spatial and time scales). A multisite evaluation will provide more certainty for model performances.

Despite the limitations of any model comparison project, they remain one of the foundational elements of climate science (Eyring *et al.*, 2016). Model intercomparisons help define common working practices amongst disparate modelling groups, identify broad strengths and weaknesses of different modelling approaches, build the knowledge and skills of participating scientists and help direct future community efforts to improve the skill and application of models.

5 | CONCLUSIONS

An international group of 45 scientists have evaluated the performance of 30 land surface models at a suburban site in Melbourne, Australia. Participating models vary in the complexity of their built geometry, hydrology and

anthropogenic representations. Ten error metrics are used with both physically-based and empirical benchmarks to assess the models performance.

Key study findings

- Compared to the earlier PILPS-Urban model comparison at the same site (Grimmond *et al.*, 2011), there is broad improvement in modelling upward short-wave radiation (SW_{up}), sensible (Q_h) and latent (Q_{le}) heat fluxes, but little/no improvement in upward long-wave radiation (LW_{up}) and momentum (Q_{tau}) fluxes.
- As in PILPS-Urban, the ensemble mean time series performs very well across all fluxes, suggesting participating models adequately span the range of uncertainty from the parameterisation of processes.
- As in PILPS-Urban, some one and two-tile urban schemes (particularly when coupled to sophisticated soil/vegetation land surface schemes), performed well across all fluxes.
- Some canyon models also perform well, indicating the integration of hydrology and vegetation into canyon models after PILPS-Urban has paid dividends.
- 'Complex' urban models are generally outperformed by others, but their overall performance is likely penalised by having simpler hydrological and vegetation approaches.
- Schemes that do not represent impervious surfaces (i.e., non-urban models), as well as urban models with simplistic hydrology/vegetation performed poorly in Q_{le}, confirming that representing both pervious and impervious surfaces is important in suburban locations.
- Detailed site information broadly improves turbulent heat fluxes but has little impact on daytime radiant fluxes.
- A two-variable out-of-sample regression outperforms most models for daytime LW_{up}, thus indicating an area for which future model development may prove beneficial.
- Many models outperform the non-linear three-variable empirical benchmarks for Q_h, with some even beating in-sample non-linear benchmarks (i.e., exceeding expected predictability using contemporaneous information). This is in stark contrast to the PLUMBER (Best *et al.*, 2015) results where no model outperforms simple SW_{down} linear regression derived from 20 non-urban sites for standard statistical metrics. It is not clear from this study if model design, model configuration, spin-up strategy and/or poorer performing benchmarks explain this.
- The empirical benchmarks may be less effective in urban locations because of anthropogenic (human behavioural) influences on fluxes, or non-contemporaneous information (e.g., memory effects of surface heat storage) being more important at urban sites, particularly at night. This implies more complex modelling techniques (i.e., land surface models rather than simple empirical models) may provide greater benefit in urban landscapes.
- Results are based on a site previously used in evaluation studies. When the details of a site are not known and not previously modelled, we should not expect such a high level of performance.

Recommendations and lessons learnt from this project:

- We recommend the use of benchmarks when evaluating models to help guide performance expectations. In this project, simple information-limited models set minimum expectations, while more complex in-sample empirical models helped indicate an upper bound for performance expectations.
- Model evaluations traditionally consider observational and modelling errors caused by parameterisation design decisions but should also explicitly consider errors caused by human factors (communication or coding errors in model or postprocessing code).
- Human errors can be reduced (but probably not eliminated) by providing participants with initial feedback and allowing resubmission prior to final analysis. We recommend the use of web-based analysis portals (e.g., modelevaluation.org) that can provide immediate feedback to participants (plots and error statistics), particularly:
  - Indicating variables that exceed expected physical limits, as well as checks on energy closure, as this helps identify model numerical errors, or errors in submitted variable's identification, units or sign.
  - Correlation of modelled vs observed short-wave radiation, as this flux varies nearly linearly with forcing, helping indicate time-of-day human errors.
- Model configuration files (e.g., parameter namelists) and model revision numbers should be submitted with model outputs to help ascertain why outputs have changed between submissions, and allow submissions to be reproducible.
- Participating in a model intercomparison project can be time-consuming. However, intercomparisons are useful for improving our understanding of model performances in general, as well as providing

opportunities to build the experience and skills of those who participate. Hence this project's methods, data and results could be used as a training tool for new modellers, in addition to providing benchmarks to test future model developments.

AFFILIATIONS

[1]Australian Research Council Centre of Excellence for Climate System Science, Climate Change Research Centre, Level 4, Mathews Building, UNSW Sydney, Sydney, New South Wales Australia

[2]Bureau of Meteorology, Sydney, New South Wales Australia

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

[4]Met Office, Exeter, UK

[5]Australian Research Council Centre of Excellence for Climate Extremes, Climate Change Research Centre, Level 4, Mathews Building, UNSW Sydney, Sydney, New South Wales Australia

[6]School of Earth, Atmosphere and Environment, Monash University, Melbourne, Australia

[7]School of Earth and Environmental Sciences, Seoul National University, Seoul, South Korea

[8]Klimaat Consulting & Innovation Inc, Guelph, Ontario Canada

[9]Met Office, University of Reading, Reading, UK

[10]European Centre for Medium-Range Weather Forecasts, Reading, UK

[11]Department of Civil and Environmental Engineering, Princeton University, Princeton, New Jersey USA

[12]School of Biological Sciences, University of Bristol, Bristol, UK

[13]CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

[14]Urban Climatology Group, Department of Geography, Ruhr-University Bochum, Bochum, Germany

[15]Department of Civil and Environmental Engineering, National University of Singapore, Singapore, Singapore

[16]Department of Meteorology and Climatology, Faculty of Geographical Sciences, University of Lodz, Lodz, Poland

[17]Department of Environment and Energy, Semyung University, Jecheon, South Korea

[18]School of Science and Engineering, Meisei University, Hino, Japan

[19]Environmental Management Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan

[20]Japan Weather Association, Tokyo, Japan

[21]Department of Atmospheric Science, Kongju National University, Gongju, Republic of Korea

[22]School of Architecture, Civil and Environmental Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[23]Department of Environment, CIEMAT, Madrid, Spain

[24]Department of Civil and Environmental Engineering, Imperial College London, London, UK

[25]Transportation, Health and Urban Design Research Lab, Faculty of Architecture, Building and Planning, University of Melbourne, Melbourne, Australia

[26]Climate and Global Dynamics Laboratory, National Center for Atmospheric Research, Boulder, Colorado USA

[27]School of Environmental Engineering, University of Seoul, Dongdaemun-gu, South Korea

[28]TECNALIA, Basque Research and Technology Alliance (BRTA), Derio, Spain

[29]Meteorology and Air Quality Section, Wageningen University, Wageningen, The Netherlands

[30]Institute for Risk and Disaster Reduction, University College London, London, UK

[31]CSIRO Environment, Commonwealth Scientific and Industrial Research Organisation, Melbourne, Victoria Australia

[32]European Centre for Medium-Range Weather Forecasts, Bonn, Germany

[33]Research Computing Center, Lomonosov, Moscow State University, Moscow, Russia

[34]A.M. Obukhov Institute of Atmospheric Physics, Moscow, Russia

[35]School of Meteorology, University of Oklahoma, Norman, Oklahoma USA

[36]Department of Geography and Environmental Sustainability, University of Oklahoma, Norman, Oklahoma USA

[37]School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, Arizona USA

ACKNOWLEDGEMENTS

The project's coordinating team is supported by UNSW Sydney, the Australian Research Council (ARC) Centre of Excellence for Climate System Science (grant CE110001028), University of Reading, the Met Office UK, the Bureau of Meteorology, Australia, the ARC Centre of Excellence for Climate Extremes (grant CE170100023) and ERC urbisphere (grant 855005). Computation support from National Computational Infrastructure (NCI) Australia. G.J. Steeneveld and A. Tsiringakis acknowledge support from the NWO VIDI grant 'The Windy City' under number 864.14.007. M. Demuzere acknowledges support from the ENLIGHT project, funded by the German Research Foundation (DFG) under grant No. 437467569. Y. Takane acknowledges support from Japan Society for the Promotion of Science (JSPS) KAKENHI Grand Number 20KK0096. Contributions by K.W. Oleson are supported by the National Center for Atmospheric Research (NCAR), sponsored by the National Science Foundation (NSF) under Cooperative Agreement No. 1852977. Computing and data storage resources for CLMU5, including the Cheyenne supercomputer (doi:10.5065/D6RX99HX), were provided by the Computational and Information Systems Laboratory (CISL) at NCAR. K. Nice acknowledges support from NHMRC/UKRI grant (1194959). J.-J. Baik acknowledges support from the National Research Foundation of Korea (NRF) under grant 2021R1A2C1007044. E. Bou-Zeid was supported by the US National Science Foundation under award number AGS 2128345 and the Army Research Office under contract W911NF2010216. Work with TERRA model performed by M. Varentsov was supported by the Russian Science Foundation, grant no. 21-17-00249. S.-H. Lee acknowledges support from the

Nuclear Safety and Security Commission (NSSC) of the Republic of Korea (No. 2105036). M. De Kauwe acknowledges support from the Natural Environment Research Council (NE/W010003/1). N. Meili and S. Fatichi acknowledge the support of the National University of Singapore through the project 'Bridging scales from below: The role of heterogeneities in the global water and carbon budgets', Award No. 22-3637-A0001. T. Sun was supported by UKRI NERC Independent Research Fellowship (NE/P018637/1 and NE/P018637/2). D.-I. Lee acknowledges support from the Korea Meteorological Administration Research and Development Program under grant KMI(KMI2021-03512).

We acknowledge the participants of the first urban model comparison project (PILPS-Urban) and all those involved in model development since then. We also acknowledge all scientists involved in collecting and providing observations used to derive benchmarks in this study. Thank you to Belinda Roux, Asiful Islam and Vinod Kumar (Bureau of Meteorology) for reviewing this manuscript. This project uses modified Copernicus Climate Change Service Information. Open access publishing facilitated by University of New South Wales, as part of the Wiley - University of New South Wales agreement via the Council of Australian University Librarians.

DATA AVAILABILITY STATEMENT

Supporting information and associated model results are available from https://urban-plumber.github.io/AU-Preston/plots/ and archived at https://doi.org/10.5281/zenodo.7388342 (Lipson *et al.*, 2022b). Observation time series data are openly available from https://doi.org/10.5281/zenodo.7104984 (Lipson *et al.*, 2022c). Benchmark time series data are available from https://doi.org/10.5281/zenodo.7330052 (Lipson and Best, 2022).

ORCID

Mathew J. Lipson https://orcid.org/0000-0001-5322-1796
Sue Grimmond https://orcid.org/0000-0002-3166-9415
Gab Abramowitz https://orcid.org/0000-0002-4205-001X
Jong-Jin Baik https://orcid.org/0000-0003-3709-0532
Lewis Blunn https://orcid.org/0000-0002-3207-5002
Elie Bou-Zeid https://orcid.org/0000-0002-6137-8109
Martin G. De Kauwe https://orcid.org/0000-0002-3399-9098
Matthias Demuzere https://orcid.org/0000-0003-3237-4077
Simone Fatichi https://orcid.org/0000-0003-1361-6659
Krzysztof Fortuniak https://orcid.org/0000-0001-7043-8751
Margaret A. Hendry https://orcid.org/0000-0003-3941-7543
Yukihiro Kikegawa https://orcid.org/0000-0002-5225-653X
Sang-Hyun Lee https://orcid.org/0000-0002-7998-9194
Gabriele Manoli https://orcid.org/0000-0002-9245-2877
Naika Meili https://orcid.org/0000-0001-6283-2134
David Meyer https://orcid.org/0000-0002-7071-7547
Kerry A. Nice https://orcid.org/0000-0001-6102-1292
Keith W. Oleson https://orcid.org/0000-0002-0057-9900
Michael Roth https://orcid.org/0000-0001-6399-3693
Andrés Simón-Moral https://orcid.org/0000-0002-2662-9750
Gert-Jan Steeneveld https://orcid.org/0000-0002-5922-8179
Ting Sun https://orcid.org/0000-0002-2486-6146

REFERENCES

Abramowitz, G. (2018) Towards improved standardisation of model evaluation using modelevaluation.org. H54A-06, 2018.

Balsamo, G., Beljaars, A., Scipal, K., Viterbo, P., van den Hurk, B., Hirschi, M. et al. (2009) A revised hydrology for the ECMWF model: verification from field site to terrestrial water storage and impact in the integrated forecast system. *Journal of Hydrometeorology*, 10, 623–643. Available from: https://doi.org/10.1175/2008JHM1068.1

Beck, H.E., Zimmermann, N.E., McVicar, T.R., Vergopolan, N., Berg, A. & Wood, E.F. (2018) Present and future Köppen-Geiger climate classification maps at 1-km resolution. *Scientific Data*, 5, 180214. Available from: https://doi.org/10.1038/sdata.2018.214

Best, M.J. & Grimmond, C.S.B. (2014) Importance of initial state and atmospheric conditions for urban land surface models' performance. *Urban Climate*, 10, 387–406. Available from: https://doi.org/10.1016/j.uclim.2013.10.006

Best, M.J. & Grimmond, C.S.B. (2015) Key conclusions of the first international urban land surface model comparison project. *Bulletin of the American Meteorological Society*, 96, 805–819. Available from: https://doi.org/10.1175/BAMS-D-14-00122.1

Best, M.J. & Grimmond, C.S.B. (2016a) Investigation of the impact of anthropogenic heat flux within an urban land surface model and PILPS-urban. *Theoretical and Applied Climatology*, 126, 51–60. Available from: https://doi.org/10.1007/s00704-015-1554-3

Best, M.J. & Grimmond, C.S.B. (2016b) Modeling the partitioning of turbulent fluxes at urban sites with varying vegetation cover. *Journal of Hydrometeorology*, 17, 2537–2553. Available from: https://doi.org/10.1175/JHM-D-15-0126.1

Best, M.J. (2005) Representing urban areas within operational numerical weather prediction models. *Boundary-Layer Meteorol*, 114, 91–109. Available from: https://doi.org/10.1007/s10546-004-4834-5

Best, M.J. (2006) Progress towards better weather forecasts for city dwellers: from short range to climate change. *Theoretical and Applied Climatology*, 84, 47–55. Available from: https://doi.org/10.1007/s00704-005-0143-2

Best, M.J., Abramowitz, G., Johnson, H.R., Pitman, A.J., Balsamo, G., Boone, A. et al. (2015) The plumbing of land surface models: benchmarking model performance. *Journal of Hydrometeorology*,

16, 1425–1442. Available from: https://doi.org/10.1175/JHM-D-14-0158.1

Best, M.J., Grimmond, C.S.B. & Villani, M.G. (2006) Evaluation of the urban tile in MOSES using surface energy balance observations. *Boundary-Layer Meteorol*, 118, 503–525. Available from: https://doi.org/10.1007/s10546-005-9025-5

Best, M.J., Pryor, M., Clark, D.B., Rooney, G.G., Essery, R.L.H., Ménard, C.B. et al. (2011) The joint UK land environment simulator (JULES), model description – part 1: energy and water fluxes. *Geoscientific Model Development*, 4, 677–699. Available from: https://doi.org/10.5194/gmd-4-677-2011

Bjorkegren, A.B., Grimmond, C.S.B., Kotthaus, S. & Malamud, B.D. (2015) CO2 emission estimation in the urban environment: measurement of the CO2 storage term. *Atmospheric Environment*, 122, 775–790. Available from: https://doi.org/10.1016/j.atmosenv.2015.10.012

Boussetta, S., Balsamo, G., Beljaars, A., Panareda, A.-A., Calvet, J.-C., Jacobs, C. et al. (2013) Natural land carbon dioxide exchanges in the ECMWF integrated forecasting system: implementation and offline validation. *Journal of Geophysical Research: Atmospheres*, 118, 5923–5946. Available from: https://doi.org/10.1002/jgrd.50488

Bowling, L. & Polcher, J. (2001) The ALMA data exchange convention. https://www.lmd.jussieu.fr/˜polcher/ALMA/

Bowling, L.C., Lettenmaier, D.P., Nijssen, B., Graham, L.P., Clark, D.B., El Maayar, M. et al. (2003) Simulation of high-latitude hydrological processes in the Torne–Kalix basin: PILPS phase 2(e): 1: experiment description and summary intercomparisons. *Global and Planetary Change*, 38, 1–30. Available from: https://doi.org/10.1016/S0921-8181(03)00003-1

Broadbent, A.M., Coutts, A.M., Nice, K.A., Demuzere, M., Krayenhoff, E.S., Tapper, N.J. et al. (2019) The air-temperature response to green/blue-infrastructure evaluation tool (TARGET v1.0): an efficient and user-friendly model of city cooling. *Geoscientific Model Development*, 12, 785–803. Available from: https://doi.org/10.5194/gmd-12-785-2019

Bueno, B., Pigeon, G., Norford, L.K., Zibouche, K. & Marchadier, C. (2012) Development and evaluation of a building energy model integrated in the TEB scheme. *Geoscientific Model Development*, 5, 433–448. Available from: https://doi.org/10.5194/gmd-5-433-2012

Chen, F. & Dudhia, J. (2001) Coupling an advanced land surface–hydrology model with the Penn State–NCAR MM5 modeling system. Part I: model implementation and sensitivity. *Monthly Weather Review*, 129, 569–585. Available from: https://doi.org/10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2

Chen, F., Kusaka, H., Bornstein, R., Ching, J., Grimmond, C.S.B., Grossman-Clarke, S. et al. (2011) The integrated WRF/urban modelling system: development, evaluation, and applications to urban environmental problems. *International Journal of Climatology*, 31, 273–288. Available from: https://doi.org/10.1002/joc.2158

Chow, W. (2017) *Eddy Covariance Data Measured at the CAP LTER Flux Tower Located in the West Phoenix, AZ Neighborhood of Maryvale from 2011-12-16 through 2012-12-31*. Environmental Data Initiative. Available from: https://doi.org/10.6073/PASTA/FED17D67583EDA16C439216CA40B0669

Chow, W.T.L., Volo, T.J., Vivoni, E.R., Jenerette, G.D. & Ruddell, B.L. (2014) Seasonal dynamics of a suburban energy balance in Phoenix, Arizona. *International Journal of Climatology*, 34, 3863–3880. Available from: https://doi.org/10.1002/joc.3947

Christen, A., Coops, N.C., Crawford, B.R., Kellett, R., Liss, K.N., Olchovski, I. et al. (2011) Validation of modeled carbon-dioxide emissions from an urban neighborhood with direct eddy-covariance measurements. *Atmospheric Environment*, 45, 6057–6069. Available from: https://doi.org/10.1016/j.atmosenv.2011.07.040

Coutts, A.M. (2006) *The Influence of Housing Density and Urban Design on the Surface Energy Balance and Local Climates of Melbourne, Australia, and the Impact of Melbourne 2030's Vision*, Thesis PhD. Melbourne, Australia: Monash University.

Coutts, A.M., Beringer, J. & Tapper, N.J. (2007a) Characteristics influencing the variability of urban CO2 fluxes in Melbourne, Australia. *Atmospheric Environment*, 41, 51–62. Available from: https://doi.org/10.1016/j.atmosenv.2006.08.030

Coutts, A.M., Beringer, J. & Tapper, N.J. (2007b) Impact of increasing urban density on local Climate: spatial and temporal variations in the surface energy balance in Melbourne, Australia. *Journal of Applied Meteorology and Climatology*, 46, 477–493. Available from: https://doi.org/10.1175/JAM2462.1

Crawford, B. & Christen, A. (2015) Spatial source attribution of measured urban eddy covariance CO2 fluxes. *Theoretical and Applied Climatology*, 119, 733–755. Available from: https://doi.org/10.1007/s00704-014-1124-0

Crawford, B., Grimmond, C.S.B. & Christen, A. (2011) Five years of carbon dioxide fluxes measurements in a highly vegetated suburban area. *Atmospheric Environment*, 45, 896–905. Available from: https://doi.org/10.1016/j.atmosenv.2010.11.017

Cucchi, M., Weedon, G.P., Amici, A., Bellouin, N., Lange, S., Müller Schmied, H. et al. (2020) WFDE5: bias-adjusted ERA5 reanalysis data for impact studies. *Earth System Science Data*, 12, 2097–2120. Available from: https://doi.org/10.5194/essd-12-2097-2020

Daniel, M., Lemonsu, A., Déqué, M., Somot, S., Alias, A. & Masson, V. (2019) Benefits of explicit urban parameterization in regional climate modeling to study climate and city interactions. *Climate Dynamics*, 52, 2745–2764. Available from: https://doi.org/10.1007/s00382-018-4289-x

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*, 14, 3835–3873. Available from: https://doi.org/10.5194/essd-14-3835-2022

Duursma, R.A. & Medlyn, B.E. (2012) MAESPA: a model to study interactions between water limitation, environmental drivers and vegetation function at tree and stand levels, with an example application to $[CO_2] \times$ drought interactions. *Geoscientific Model Development*, 5, 919–940. Available from: https://doi.org/10.5194/gmd-5-919-2012

ECMWF. (2020) IFS Documentation CY47R1 – Part IV: Physical Processes. https://doi.org/10.21957/CPMKQVHJA

Eyring, V., Bony, S., Meehl, G.A., Senior, C.A., Stevens, B., Stouffer, R.J. et al. (2016) Overview of the coupled model Intercomparison project phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9, 1937–1958. Available from: https://doi.org/10.5194/gmd-9-1937-2016

Fatichi, S., Ivanov, V.Y. & Caporali, E. (2012) A mechanistic ecohydrological model to investigate complex interactions in cold and warm water-controlled environments: 1. Theoretical framework and plot-scale analysis. *Journal of Advances in Modeling Earth Systems*, 4, M05002. Available from: https://doi.org/10.1029/2011MS000086

Flanner, M.G. (2009) Integrating anthropogenic heat flux with global climate models. *Geophysical Research Letters*, 36, L02801. Available from: https://doi.org/10.1029/2008GL036465

Fortuniak, K. (2003) A slab surface energy balance model (SUEB) and its application to the study on the role of roughness length in forming an urban heat Island. *Acta Universitatis Wratislaviensis*, 2542, 368–377.

Fortuniak, K., Kłysik, K. & Siedlecki, M. (2006) New measurements of the energy balance components in Łódź. In: *Preprints, Sixth International Conference on Urban Climate: 12–16 June, 2006, Göteborg, Sweden, Sixth International Conference on Urban Climate*. Göteborg, Sweden: International Association for Urban Climate, pp. 64–67.

Fortuniak, K., Pawlak, W. & Siedlecki, M. (2013) Integral turbulence statistics over a central European city Centre. *Boundary Layer Meteorology; Dordrecht*, 146, 257–276. Available from: https://doi.org/10.1007/s10546-012-9762-1

Garbero, V., Milelli, M., Bucchignani, E., Mercogliano, P., Varentsov, M., Rozinkina, I. et al. (2021) Evaluating the urban canopy scheme TERRA_URB in the COSMO model for selected European cities. *Atmosphere*, 12, 237. Available from: https://doi.org/10.3390/atmos12020237

Garuma, G.F. (2018) Review of urban surface parameterizations for numerical climate models. *Urban Climate*, 24, 830–851. Available from: https://doi.org/10.1016/j.uclim.2017.10.006

Goret, M., Masson, V., Schoetter, R. & Moine, M.-P. (2019) Inclusion of CO2 flux modelling in an urban canopy layer model and an evaluation over an old European city Centre. *Atmospheric Environment: X*, 3, 100042. Available from: https://doi.org/10.1016/j.aeaoa.2019.100042

Grimmond, C.S.B., Best, M., Barlow, J., Arnfield, A.J., Baik, J.-J., Baklanov, A. et al. (2009) Urban surface energy balance models: model characteristics and methodology for a comparison study. In: Baklanov, A., Sue, G., Alexander, M. & Athanassiadou, M. (Eds.) *Meteorological and Air Quality Models for Urban Areas*. Berlin Heidelberg: Springer, pp. 97–123.

Grimmond, C.S.B., Blackett, M., Best, M.J., Baik, J.-J., Belcher, S.E., 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. Available from: https://doi.org/10.1002/joc.2227

Grimmond, C.S.B., Blackett, M., Best, M.J., Barlow, J., Baik, J.-J., Belcher, S.E. 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. Available from: https://doi.org/10.1175/2010JAMC2354.1

Grimmond, C.S.B., Cleugh, H.A. & Oke, T.R. (1991) An objective urban heat storage model and its comparison with other schemes. *Atmospheric Environment. Part B. Urban Atmosphere*, 25, 311–326. Available from: https://doi.org/10.1016/0957-1272(91)90003-W

Hamdi, R. & Masson, V. (2008) Inclusion of a drag approach in the town energy balance (TEB) scheme: offline 1D evaluation in a street canyon. *Journal of Applied Meteorology and Climatology*, 47, 2627–2644. Available from: https://doi.org/10.1175/2008JAMC1865.1

Haughton, N., Abramowitz, G. & Pitman, A.J. (2017) On the predictability of land surface fluxes from meteorological variables. *Geoscientific Model Development Discussion*, 1–27, 2017. Available from: https://doi.org/10.5194/gmd-2017-153

Haughton, N., Abramowitz, G., Pitman, A.J., Or, D., Best, M.J., Johnson, H.R. et al. (2016) The plumbing of land surface models: is poor performance a result of methodology or data quality? *Journal of Hydrometeorology*, 17, 1705–1723. Available from: https://doi.org/10.1175/JHM-D-15-0171.1

Henderson-Sellers, A., McGuffie, K. & Pitman, A.J. (1996) The project for Intercomparison of land-surface parametrization schemes (PILPS): 1992 to 1995. *Climate Dynamics*, 12, 849–859. Available from: https://doi.org/10.1007/s003820050147

Henderson-Sellers, A., Pitman, A.J., Love, P.K., Irannejad, P. & Chen, T.H. (1995) The project for Intercomparison of land surface parameterization schemes (PILPS): phases 2 and 3*. *Bulletin of the American Meteorological Society*, 76, 489–504. Available from: https://doi.org/10.1175/1520-0477(1995)076<0489:TPFIOL>2.0.CO;2

Hengl, T. (2018a) Clay content in % (kg/kg) at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution (v0.2). https://doi.org/10.5281/ZENODO.2525663

Hengl, T. (2018b) Sand content in % (kg/kg) at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution (v0.2). https://doi.org/10.5281/ZENODO.2525662

Hengl, T. (2018c) Soil bulk density (fine earth) 10 × kg/m-cubic at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution (v0.2). https://doi.org/10.5281/ZENODO.2525665

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J. et al. (2018) ERA5 hourly data on single levels from 1979 to present. *Copernicus Climate Change Service (c3s) Climate Data Store (cds)*, 10. https://doi.org/10.24381/cds.adbb2d47

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J. et al. (2020) The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146, 1999–2049. Available from: https://doi.org/10.1002/qj.3803

Hirano, T., Sugawara, H., Murayama, S. & Kondo, H. (2015) Diurnal variation of CO_2 flux in an urban area of Tokyo. *Scientific Online Letters On The Atmosphere*, 11, 100–103. Available from: https://doi.org/10.2151/sola.2015-024

Hogan, R.J. (2019a) An exponential model of urban geometry for use in radiative transfer applications. *Boundary-Layer Meteorol*, 170, 357–372. Available from: https://doi.org/10.1007/s10546-018-0409-8

Hogan, R.J. (2019b) Flexible treatment of radiative transfer in complex urban canopies for use in weather and Climate models. *Boundary-Layer Meteorol*, 173, 53–78. Available from: https://doi.org/10.1007/s10546-019-00457-0

Hollinger, D.Y. & Richardson, A.D. (2005) Uncertainty in eddy covariance measurements and its application to physiological models. *Tree Physiology*, 25, 873–885. Available from: https://doi.org/10.1093/treephys/25.7.873

Hong, J., Lee, K. & Hong, J.-W. (2020) Observational data of Ochang and Jungnang in Korea. https://doi.org/10.22647/EAPL-OC_JN2021

Hong, J.-W., Hong, J., Chun, J., Lee, Y.H., Chang, L.-S., Lee, J.-B. et al. (2019) Comparative assessment of net CO2 exchange across an urbanization gradient in Korea based on eddy covariance measurements. *Carbon Balance and Management*, 14, 13. Available from: https://doi.org/10.1186/s13021-019-0128-6

Hong, S.-O., Kim, J., Byun, Y.-H., Hong, J., Hong, J.-W., Lee, K. et al. (2023) Intra-urban variations of the CO2 fluxes at the surface-atmosphere interface in the seoul metropolitan area. *Asia-Pacific Journal of Atmospheric Sciences*, 59(4), 417–431. Available from: https://doi.org/10.1007/s13143-023-00324-6

Ishidoya, S., Sugawara, H., Terao, Y., Kaneyasu, N., Aoki, N., Tsuboi, K. et al. (2020) O2 : CO2 exchange ratio for net turbulent flux observed in an urban area of Tokyo, Japan, and its application to an evaluation of anthropogenic CO2 emissions. *Atmospheric Chemistry and Physics*, 20, 5293–5308. Available from: https://doi.org/10.5194/acp-20-5293-2020

Jackson, E.K., Roberts, W., Nelsen, B., Williams, G.P., Nelson, E.J. & Ames, D.P. (2019) Introductory overview: error metrics for hydrologic modelling – a review of common practices and an open source library to facilitate use and adoption. *Environmental Modelling & Software*, 119, 32–48. Available from: https://doi.org/10.1016/j.envsoft.2019.05.001

Jackson, T.L., Feddema, J.J., Oleson, K.W., Bonan, G.B. & Bauer, J.T. (2010) Parameterization of urban characteristics for global Climate modeling. *Annals of the Association of American Geographers*, 100, 848–865. Available from: https://doi.org/10.1080/00045608.2010.497328

Järvi, L., Grimmond, C.S.B. & Christen, A. (2011) The surface urban energy and water balance scheme (SUEWS): evaluation in Los Angeles and Vancouver. *Journal of Hydrology*, 411, 219–237. Available from: https://doi.org/10.1016/j.jhydrol.2011.10.001

Järvi, L., Rannik, Ü., Kokkonen, T.V., Kurppa, M., Karppinen, A., Kouznetsov, R.D. et al. (2018) Uncertainty of eddy covariance flux measurements over an urban area based on two towers. *Atmospheric Measurement Techniques*, 11, 5421–5438. Available from: https://doi.org/10.5194/amt-11-5421-2018

Kanda, M., Kawai, T., Kanega, M., Moriwaki, R., Narita, K. & Hagishima, A. (2005) A simple energy balance model for regular building arrays. *Boundary-Layer Meteorol*, 116, 423–443. Available from: https://doi.org/10.1007/s10546-004-7956-x

Karsisto, P., Fortelius, C., Demuzere, M., Grimmond, C.S.B., Oleson, K.W., Kouznetsov, R. et al. (2016) Seasonal surface urban energy balance and wintertime stability simulated using three land-surface models in the high-latitude city Helsinki. *Quarterly Journal of the Royal Meteorological Society*, 142, 401–417. Available from: https://doi.org/10.1002/qj.2659

Kikegawa, Y., Genchi, Y., Kondo, H. & Hanaki, K. (2006) Impacts of city-block-scale countermeasures against urban heat-Island phenomena upon a building's energy-consumption for air-conditioning. *Applied Energy*, 83, 649–668. Available from: https://doi.org/10.1016/j.apenergy.2005.06.001

Kikegawa, Y., Genchi, Y., Yoshikado, H. & Kondo, H. (2003) Development of a numerical simulation system toward comprehensive assessments of urban warming countermeasures including their impacts upon the urban buildings' energy-demands. *Applied Energy*, 76, 449–466. Available from: https://doi.org/10.1016/S0306-2619(03)00009-6

Kikegawa, Y., Tanaka, A., Ohashi, Y., Ihara, T. & Shigeta, Y. (2014) Observed and simulated sensitivities of summertime urban surface air temperatures to anthropogenic heat in downtown areas of two Japanese major cities, Tokyo and Osaka. *Theoretical and Applied Climatology*, 117, 175–193. Available from: https://doi.org/10.1007/s00704-013-0996-8

Kondo, H. & Liu, F.-H. (1998) A study on the urban thermal environment obtained through one-dimensional urban canopy model. *Journal of Japan Society for Atmospheric Environment / Taiki Kankyo Gakkaishi*, 33, 179–192. Available from: https://doi.org/10.11298/taiki1995.33.3&uscore;179

Kondo, H., Genchi, Y., Kikegawa, Y., Ohashi, Y., Yoshikado, H. & Komiyama, H. (2005) Development of a multi-layer urban canopy model for the analysis of energy consumption in a big City: structure of the urban canopy model and its basic performance. *Boundary-Layer Meteorol*, 116, 395–421. Available from: https://doi.org/10.1007/s10546-005-0905-5

Kondo, H., Inagaki, A. & Kanda, M. (2015) A new parametrization of mixing length in an urban canopy derived from a large-Eddy simulation database for Tokyo. *Boundary-Layer Meteorol*, 156, 131–144. Available from: https://doi.org/10.1007/s10546-015-0019-7

Koster, R.D., Guo, Z., Yang, R., Dirmeyer, P.A., Mitchell, K. & Puma, M.J. (2009) On the nature of soil moisture in land surface models. *Journal of Climate*, 22, 4322–4335. Available from: https://doi.org/10.1175/2009JCLI2832.1

Kotthaus, S. & Grimmond, C.S.B. (2014a) Energy exchange in a dense urban environment – part I: temporal variability of long-term observations in Central London. *Urban Climate*, 10, 261–280. Available from: https://doi.org/10.1016/j.uclim.2013.10.002

Kotthaus, S. & Grimmond, C.S.B. (2014b) Energy exchange in a dense urban environment – part II: impact of spatial heterogeneity of the surface. *Urban Climate*, 10, 281–307. Available from: https://doi.org/10.1016/j.uclim.2013.10.001

Kowalczyk, E., Stevens, L., Law, R., Dix, M., Wang, Y., Harman, I. et al. (2013) The land surface model component of ACCESS: description and impact on the simulated surface climatology. *Australian Meteorological and Oceanographic Journal*, 63, 65–82.

Kowalczyk, E.A., Wang, Y.P., Law, R.M., Davies, H.L., McGregor, J.L. & Abramowitz, G.S. (2006) The CSIRO atmosphere biosphere land exchange (CABLE) model for use in climate models and as an offline model. https://doi.org/10.4225/08/58615C6A9A51D

Krayenhoff, E.S. & Voogt, J.A. (2007) A microscale three-dimensional urban energy balance model for studying surface temperatures. *Boundary-Layer Meteorol*, 123, 433–461. Available from: https://doi.org/10.1007/s10546-006-9153-6

Kusaka, H., Kondo, H., Kikegawa, Y. & Kimura, F. (2001) A simple single-layer urban canopy model for atmospheric models: comparison with multi-layer and slab models. *Boundary-Layer Meteorology*, 101, 329–358. Available from: https://doi.org/10.1023/A:1019207923078

Le Moigne, P., Albergel, C., Boone, A., Belamari, S., Decharme, B., Dumont, M. et al. (2018) SURFEX v8.1 Scientific Documentation.

Lee, D.-I. & Lee, S.-H. (2020) The microscale urban surface energy (MUSE) model for real urban application. *Atmosphere*, 11, 1347. Available from: https://doi.org/10.3390/atmos11121347

Lee, S.-H. & Park, S.-U. (2008) A vegetated urban canopy model for meteorological and environmental modelling. *Boundary-Layer Meteorol*, 126, 73–102. Available from: https://doi.org/10.1007/s10546-007-9221-6

Lee, S.-H. (2011) Further development of the vegetated urban canopy model including a grass-covered surface parametrization and photosynthesis effects. *Boundary-Layer Meteorol*, 140, 315–342. Available from: https://doi.org/10.1007/s10546-011-9603-7

Lee, S.-H., Lee, H., Park, S.-B., Woo, J.-W., Lee, D.-I. & Baik, J.-J. (2016) Impacts of in-canyon vegetation and canyon aspect ratio on the thermal environment of street canyons: numerical investigation using a coupled WRF-VUCM model. *Quarterly Journal of the Royal Meteorological Society*, 142, 2562–2578. Available from: https://doi.org/10.1002/qj.2847

Lemonsu, A., Masson, V., Shashua-Bar, L., Erell, E. & Pearlmutter, D. (2012) Inclusion of vegetation in the town energy balance model for modelling urban green areas. *Geoscientific Model*

Development, 5, 1377–1393. Available from: https://doi.org/10.5194/gmd-5-1377-2012

Lipson, M.J., Thatcher, M., Hart, M.A. & Pitman, A. (2018) A building energy demand and urban land surface model. *Quarterly Journal of the Royal Meteorological Society*, 144, 1572–1590. Available from: https://doi.org/10.1002/qj.3317

Lipson, M. & Best, M. (2022) Benchmarks for the urban-PLUMBER model evaluation project phase 1 (AU-Preston). https://doi.org/10.5281/zenodo.7330052

Lipson, M., Grimmond, S., Best, M., Abramowitz, G., Coutts, A., Tapper, N. et al. (2022b) Associated results of phase 1 of the urban-PLUMBER model evaluation project. https://doi.org/10.5281/zenodo.7388342

Lipson, M., Grimmond, S., Best, M., Abramowitz, G., Kauwe, M.D., Tsiringakis, A. et al. (2020) Modelling protocol for the urban-PLUMBER model evaluation project. https://doi.org/10.5281/zenodo.6363850

Lipson, M., Grimmond, S., Best, M., Chow, W., Christen, A., Chrysoulakis, N. et al. (2022c) Data for "Harmonized gap-filled dataset from 20 urban flux tower sites" for the Urban-PLUMBER project. https://doi.org/10.5281/zenodo.7104984

Lipson, M., Grimmond, S., Best, M., Chow, W.T.L., Christen, A., Chrysoulakis, N. et al. (2022a) Harmonized gap-filled datasets from 20 urban flux tower sites. *Earth System Science Data*, 14, 5157–5178. Available from: https://doi.org/10.5194/essd-14-5157-2022

Lipson, M.J., Hart, M.A. & Thatcher, M. (2017) Efficiently modelling urban heat storage: an interface conduction scheme in an urban land surface model (aTEB v2.0). *Geoscientific Model Development*, 10, 991–1007. Available from: https://doi.org/10.5194/gmd-10-991-2017

Liu, Y., Chen, F., Warner, T. & Basara, J. (2006) Verification of a mesoscale data-assimilation and forecasting system for the Oklahoma City area during the joint urban 2003 field project. *Journal of Applied Meteorology and Climatology*, 45, 912–929. Available from: https://doi.org/10.1175/JAM2383.1

Manabe, S. (1969) Climate and the ocean circulation: I. The atmospheric circulation and the hydrology of the EARTH'S surface. *Monthly Weather Review*, 97, 739–774. Available from: https://doi.org/10.1175/1520-0493(1969)097<0739:CATOC>2.3.CO;2

Martens, B., Schumacher, D.L., Wouters, H., Muñoz-Sabater, J., Verhoest, N.E.C. & Miralles, D.G. (2020) Evaluating the land-surface energy partitioning in ERA5. *Geoscientific Model Development*, 13, 4159–4181. Available from: https://doi.org/10.5194/gmd-13-4159-2020

Martilli, A., Clappier, A. & Rotach, M.W. (2002) An urban surface exchange parameterisation for mesoscale models. *Boundary-Layer Meteorology*, 104, 261–304. Available from: https://doi.org/10.1023/A:1016099921195

Martilli, A., Santiago, J.L. & Salamanca, F. (2015) On the representation of urban heterogeneities in mesoscale models. *Environmental Fluid Mechanics*, 15, 305–328. Available from: https://doi.org/10.1007/s10652-013-9321-4

Masson, V. (2000) A physically-based scheme forthe urban energy budget in atmospheric models. *Boundary-Layer Meteorology*, 94, 357–397. Available from: https://doi.org/10.1023/A:1002463829265

Masson, V., Gomes, L., Pigeon, G., Liousse, C., Pont, V., Lagouarde, J.-P. et al. (2008) The canopy and aerosol particles interactions in TOulouse urban layer (CAPITOUL) experiment. *Meteorology and Atmospheric Physics*, 102, 135–157. Available from: https://doi.org/10.1007/s00703-008-0289-4

Masson, V., Le Moigne, P., Martin, E., Faroux, S., Alias, A., Alkama, R. et al. (2013) The SURFEXv7.2 land and ocean surface platform for coupled or offline simulation of earth surface variables and fluxes. *Geoscientific Model Development*, 6, 929–960. Available from: https://doi.org/10.5194/gmd-6-929-2013

Masson, V., Lemonsu, A., Hidalgo, J. & Voogt, J. (2020) Urban climates and Climate change. *Annual Review of Environment and Resources*, 45, 411–444. Available from: https://doi.org/10.1146/annurev-environ-012320-083623

Masson, V., Lemonsu, A., Pigeon, G., Schoetter, R., de Munck, C., Bueno, B. et al. (2021) *The Town Energy Balance (TEB) model*. Available from: https://doi.org/10.5281/zenodo.5104731

McGregor, J.L. & Dix, M.R. (2008) An Updated Description of the Conformal-Cubic Atmospheric Model. In: Hamilton, K. & Ohfuchi, W. (Eds.) *High resolution numerical modelling of the atmosphere and Ocean*. New York, NY: Springer, pp. 51–75. Available from: https://doi.org/10.1007/978-0-387-49791-4_4

McNorton, J.R., Arduini, G., Bousserez, N., Agustí-Panareda, A., Balsamo, G., Boussetta, S. et al. (2021) An urban scheme for the ECMWF integrated forecasting system: single-column and global offline application. *Journal of Advances in Modeling Earth Systems*, 13, e2020MS002375. Available from: https://doi.org/10.1029/2020MS002375

Meili, N., Manoli, G., Burlando, P., Bou-Zeid, E., Chow, W.T.L., Coutts, A.M. et al. (2020) An urban ecohydrological model to quantify the effect of vegetation on urban climate and hydrology (UT&C v1.0). *Geoscientific Model Development*, 13, 335–362. Available from: https://doi.org/10.5194/gmd-13-335-2020

Meili, N., Manoli, G., Burlando, P., Carmeliet, J., Chow, W.T.L., Coutts, A.M. et al. (2021) Tree effects on urban microclimate: diurnal, seasonal, and climatic temperature differences explained by separating radiation, evapotranspiration, and roughness effects. *Urban Forestry & Urban Greening*, 58, 126970. Available from: https://doi.org/10.1016/j.ufug.2020.126970

Menard, C.B., Essery, R., Krinner, G., Arduini, G., Bartlett, P., Boone, A. et al. (2021) Scientific and human errors in a snow model Intercomparison. *Bulletin of the American Meteorological Society*, 102, E61–E79. Available from: https://doi.org/10.1175/BAMS-D-19-0329.1

Menzer, O. & McFadden, J.P. (2017) Statistical partitioning of a three-year time series of direct urban net CO2 flux measurements into biogenic and anthropogenic components. *Atmospheric Environment*, 170, 319–333. Available from: https://doi.org/10.1016/j.atmosenv.2017.09.049

Meyer, D. & Raustad, R. (2020) MinimalDX. https://doi.org/10.5281/zenodo.3892452

Meyer, D., Schoetter, R., Masson, V. & Grimmond, S. (2020b) Enhanced software and platform for the town energy balance (TEB) model. *Journal of Open Source Software*, 5, 2008. Available from: https://doi.org/10.21105/joss.02008

Meyer, D., Schoetter, R., Riechert, M., Verrelle, A., Tewari, M., Dudhia, J. et al. (2020a) WRF-TEB: implementation and evaluation of the coupled weather research and forecasting (WRF) and town energy balance (TEB) Model. *Journal of Advances in Modeling Earth Systems*, 12, e2019MS001961. Available from: https://doi.org/10.1029/2019MS001961

Nazarian, N., Lipson, M. & Norford, L.K. (2023) Chapter 4 – Multiscale modeling techniques to document urban climate change. In: Paolini, R. & Santamouris, M. (Eds.) *Urban Climate*

Change and Heat Islands. Netherlands: Elsevier, pp. 123–164. Available from: https://doi.org/10.1016/B978-0-12-818977-1.00004-1

Nice, K.A., Coutts, A.M. & Tapper, N.J. (2018) Development of the VTUF-3D v1.0 urban micro-climate model to support assessment of urban vegetation influences on human thermal comfort. *Urban Climate*, 24, 1052–1076. Available from: https://doi.org/10.1016/j.uclim.2017.12.008

Nordbo, A., Järvi, L., Haapanala, S., Moilanen, J. & Vesala, T. (2013) Intra-City variation in urban morphology and turbulence structure in Helsinki, Finland. *Boundary-Layer Meteorol*, 146, 469–496. Available from: https://doi.org/10.1007/s10546-012-9773-y

Oleson, K.W. & Feddema, J. (2020) Parameterization and surface data improvements and new capabilities for the community land model urban (CLMU). *Journal of Advances in Modeling Earth Systems*, 12, e2018MS001586. Available from: https://doi.org/10.1029/2018MS001586

Oleson, K.W., Anderson, G.B., Jones, B., McGinnis, S.A. & Sanderson, B. (2018) Avoided climate impacts of urban and rural heat and cold waves over the U.S. using large climate model ensembles for RCP8.5 and RCP4.5. *Climatic Change*, 146, 377–392. Available from: https://doi.org/10.1007/s10584-015-1504-1

Oleson, K.W., Bonan, G.B., Feddema, J.J., Vertenstein, M. & Kluzek, E. (2010) *Technical Description of an Urban Parameterization for the Community Land Model (CLMU)*. Boulder, Colarado: National Center for Atmospheric Research.

Pawlak, W., Fortuniak, K. & Siedlecki, M. (2011) Carbon dioxide flux in the Centre of Łódź, Poland—analysis of a 2-year eddy covariance measurement data set. *International Journal of Climatology*, 31, 232–243. Available from: https://doi.org/10.1002/joc.2247

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. et al. (2011) Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12, 2825–2830.

Perkins, S.E., Pitman, A.J., Holbrook, N.J. & McAneney, J. (2007) Evaluation of the AR4 Climate Models' simulated daily maximum temperature, minimum temperature, and precipitation over Australia using probability density functions. *Journal of Climate*, 20, 4356–4376. Available from: https://doi.org/10.1175/JCLI4253.1

Peters, E.B., Hiller, R.V. & McFadden, J.P. (2011) Seasonal contributions of vegetation types to suburban evapotranspiration. *Journal of Geophysical Research: Biogeosciences*, 116, G01003. Available from: https://doi.org/10.1029/2010JG001463

Pitman, A.J. (2003) The evolution of, and revolution in, land surface schemes designed for climate models. *International Journal of Climatology*, 23, 479–510. Available from: https://doi.org/10.1002/joc.893

Porson, A., Clark, P.A., Harman, I.N., Best, M.J. & Belcher, S.E. (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. Available from: https://doi.org/10.1002/qj.668

Redon, E.C., Lemonsu, A., Masson, V., Morille, B. & Musy, M. (2017) Implementation of street trees within the solar radiative exchange parameterization of TEB in SURFEX v8.0. *Geoscientific Model Development*, 10, 385–411. Available from: https://doi.org/10.5194/gmd-10-385-2017

Richardson, A.D., Hollinger, D.Y., Burba, G.G., Davis, K.J., Flanagan, L.B., Katul, G.G. et al. (2006) A multi-site analysis of random error in tower-based measurements of carbon and energy fluxes. *Agricultural and Forest Meteorology*, 136, 1–18. Available from: https://doi.org/10.1016/j.agrformet.2006.01.007

Rockel, B., Will, A. & Hense, A. (2008) The regional Climate model COSMO-CLM (CCLM). *Meteorologische Zeitschrift*, 17, 347–348. Available from: https://doi.org/10.1127/0941-2948/2008/0309

Roth, M., Jansson, C. & Velasco, E. (2017) Multi-year energy balance and carbon dioxide fluxes over a residential neighbourhood in a tropical city. *International Journal of Climatology*, 37, 2679–2698. Available from: https://doi.org/10.1002/joc.4873

Ryu, Y.-H., Baik, J.-J. & Lee, S.-H. (2011) A new single-layer urban canopy model for use in mesoscale atmospheric models. *Journal of Applied Meteorology and Climatology*, 50, 1773–1794. Available from: https://doi.org/10.1175/2011JAMC2665.1

Sabot, M.E.B., De Kauwe, M.G., Pitman, A.J., Medlyn, B.E., Verhoef, A., Ukkola, A.M. et al. (2020) Plant profit maximization improves predictions of European forest responses to drought. *New Phytologist*, 226, 1638–1655. Available from: https://doi.org/10.1111/nph.16376

Sailor, D.J. (2011) A review of methods for estimating anthropogenic heat and moisture emissions in the urban environment. *International Journal of Climatology*, 31, 189–199. Available from: https://doi.org/10.1002/joc.2106

Salamanca, F., Krayenhoff, E.S. & Martilli, A. (2009) On the derivation of material thermal properties representative of heterogeneous urban neighborhoods. *Journal of Applied Meteorology & Climatology*, 48, 1725–1732. Available from: https://doi.org/10.1175/2009JAMC2176.1

Schmid, H.P., Cleugh, H.A., Grimmond, C.S.B. & Oke, T.R. (1991) Spatial variability of energy fluxes in suburban terrain. *Boundary-Layer Meteorol*, 54, 249–276. Available from: https://doi.org/10.1007/BF00183956

Schoetter, R., Masson, V., Bourgeois, A., Pellegrino, M. & Lévy, J.-P. (2017) Parametrisation of the variety of human behaviour related to building energy consumption in the town energy balance (SURFEX-TEB v. 8.2). *Geoscientific Model Development*, 10, 2801–2831. Available from: https://doi.org/10.5194/gmd-10-2801-2017

Schulz, J.-P., Vogel, G., Becker, C., Kothe, S., Rummel, U. & Ahrens, B. (2016) Evaluation of the ground heat flux simulated by a multi-layer land surface scheme using high-quality observations at grass land and bare soil. *Meteorologische Zeitschrift*, 607–620, 607–620. Available from: https://doi.org/10.1127/metz/2016/0537

Sharma, A., Wuebbles, D.J. & Kotamarthi, R. (2021) The need for urban-resolving Climate modeling across scales. *AGU Advances*, 2, e2020AV000271. Available from: https://doi.org/10.1029/2020AV000271

Simón-Moral, A., Santiago, J.L. & Martilli, A. (2017) Effects of unstable thermal stratification on vertical fluxes of heat and momentum in urban areas. *Boundary-Layer Meteorol*, 163, 103–121. Available from: https://doi.org/10.1007/s10546-016-0211-4

Slater, A.G., Schlosser, C.A., Desborough, C.E., Pitman, A.J., Henderson-Sellers, A., Robock, A. et al. (2001) The representation of snow in land surface schemes: results from PILPS 2(d). *Journal of Hydrometeorology*, 2, 7–25. Available from: https://doi.org/10.1175/1525-7541(2001)002<0007:TROSIL>2.0.CO;2

Stagakis, S., Chrysoulakis, N., Spyridakis, N., Feigenwinter, C. & Vogt, R. (2019) Eddy covariance measurements and source partitioning of CO2 emissions in an urban environment: application for Heraklion Greece. *Atmospheric Environment*, 201, 278–292. Available from: https://doi.org/10.1016/j.atmosenv.2019.01.009

Stavropulos-Laffaille, X., Chancibault, K., Andrieu, H., Lemonsu, A., Calmet, I., Keravec, P. et al. (2021) Coupling detailed urban energy and water budgets with TEB-hydro model: towards an assessment tool for nature based solution performances. *Urban Climate*, 39, 100925. Available from: https://doi.org/10.1016/j.uclim.2021.100925

Stavropulos-Laffaille, X., Chancibault, K., Brun, J.-M., Lemonsu, A., Masson, V., Boone, A. et al. (2018) Improvements to the hydrological processes of the town energy balance model (TEB-veg, SURFEX v7.3) for urban modelling and impact assessment. *Geoscientific Model Development*, 11, 4175–4194. Available from: https://doi.org/10.5194/gmd-11-4175-2018

Steeneveld, G.-J., van der Horst, S. & Heusinkveld, B. (2020) Observing the surface radiation and energy balance, carbon dioxide and methane fluxes over the city Centre of Amsterdam. *Copernicus Meetings*, EGU2020-1547. Available from: https://doi.org/10.5194/egusphere-egu2020-1547

Stewart, I.D. & Oke, T.R. (2012) Local Climate zones for urban temperature studies. *Bulletin of the American Meteorological Society*, 93, 1879–1900. Available from: https://doi.org/10.1175/BAMS-D-11-00019.1

Stretton, M.A., Morrison, W., Hogan, R.J. & Grimmond, S. (2022) Evaluation of the SPARTACUS-urban radiation model for vertically resolved shortwave radiation in urban areas. *Boundary-Layer Meteorol*, 184, 301–331. Available from: https://doi.org/10.1007/s10546-022-00706-9

Takane, Y., Nakajima, K. & Kikegawa, Y. (2022) Urban climate changes during the COVID-19 pandemic: integration of urban-building-energy model with social big data. *npj Climate and Atmospheric Science*, 5, 1–10. Available from: https://doi.org/10.1038/s41612-022-00268-0

Taylor, K.E. (2001) Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research*, 106, 7183–7192. Available from: https://doi.org/10.1029/2000JD900719

Thatcher, M. & Hurley, P. (2012) Simulating Australian urban climate in a mesoscale atmospheric numerical model. *Boundary-Layer Meteorol*, 142, 149–175. Available from: https://doi.org/10.1007/s10546-011-9663-8

Tremback, C.J. & Kessler, R. (1985) A surface temperature and moisture parameterization for use in mesoscale numerical models, NTRS Author Affiliations: Colorado State Univ., Colorado State UniversityNTRS Document ID: 19870024450NTRS Research Center: Legacy CDMS (CDMS).

Varentsov, M., Samsonov, T. & Demuzere, M. (2020) Impact of urban canopy parameters on a Megacity's modelled thermal environment. *Atmosphere*, 11, 1349. Available from: https://doi.org/10.3390/atmos11121349

Velasco, E., Perrusquia, R., Jiménez, E., Hernández, F., Camacho, P., Rodríguez, S. et al. (2014) Sources and sinks of carbon dioxide in a neighborhood of Mexico City. *Atmospheric Environment*, 97, 226–238. Available from: https://doi.org/10.1016/j.atmosenv.2014.08.018

Velasco, E., Pressley, S., Grivicke, R., Allwine, E., Molina, L.T. & Lamb, B. (2011) Energy balance in urban Mexico City: observation and parameterization during the MILAGRO/MCMA-2006 field campaign. *Theoretical and Applied Climatology*, 103, 501–517. Available from: https://doi.org/10.1007/s00704-010-0314-7

Wang, C., Wang, Z.-H. & Ryu, Y.-H. (2021) A single-layer urban canopy model with transmissive radiation exchange between trees and street canyons. *Building and Environment*, 191, 107593. Available from: https://doi.org/10.1016/j.buildenv.2021.107593

Wang, Y.P., Kowalczyk, E., Leuning, R., Abramowitz, G., Raupach, M.R., Pak, B. et al. (2011a) Diagnosing errors in a land surface model (CABLE) in the time and frequency domains. *Journal of Geophysical Research: Biogeosciences*, 116, G01034. Available from: https://doi.org/10.1029/2010JG001385

Wang, Z.-H. (2014) Monte Carlo simulations of radiative heat exchange in a street canyon with trees. *Solar Energy*, 110, 704–713. Available from: https://doi.org/10.1016/j.solener.2014.10.012

Wang, Z.-H., Bou-Zeid, E. & Smith, J.A. (2011b) A spatially-analytical scheme for surface temperatures and conductive heat fluxes in urban canopy models. *Boundary-Layer Meteorol*, 138, 171–193. Available from: https://doi.org/10.1007/s10546-010-9552-6

Wang, Z.-H., Bou-Zeid, E. & Smith, J.A. (2013) A coupled energy transport and hydrological model for urban canopies evaluated using a wireless sensor network. *Quarterly Journal of the Royal Meteorological Society*, 139, 1643–1657. Available from: https://doi.org/10.1002/qj.2032

Ward, H.C., Evans, J.G. & Grimmond, C.S.B. (2013) Multi-season eddy covariance observations of energy, water and carbon fluxes over a suburban area in Swindon UK. *Atmospheric Chemistry and Physics*, 13, 4645–4666. Available from: https://doi.org/10.5194/acp-13-4645-2013

Ward, H.C., Kotthaus, S., Järvi, L. & Grimmond, C.S.B. (2016) Surface urban energy and water balance scheme (SUEWS): development and evaluation at two UK sites. *Urban Climate*, 18, 1–32. Available from: https://doi.org/10.1016/j.uclim.2016.05.001

Willmott, C.J. & Matsuura, K. (2005) Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, 30, 79–82. Available from: https://doi.org/10.3354/cr030079

Wouters, H., Demuzere, M., Blahak, U., Fortuniak, K., Maiheu, B., Camps, J. et al. (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, 3027–3054. Available from: https://doi.org/10.5194/gmd-9-3027-2016

Wouters, H., Demuzere, M., Ridder, K.D. & van Lipzig, N.P.M. (2015) The impact of impervious water-storage parametrization on urban climate modelling. *Urban Climate*, 11, 24–50. Available from: https://doi.org/10.1016/j.uclim.2014.11.005

Yang, J., Wang, Z.-H., Chen, F., Miao, S., Tewari, M., Voogt, J.A. et al. (2015) Enhancing hydrologic modelling in the coupled weather research and forecasting–urban modelling system. *Boundary-Layer Meteorol*, 155, 87–109. Available from: https://doi.org/10.1007/s10546-014-9991-6

Yang, Z.-L., Dickinson, R.E., Henderson-Sellers, A. & Pitman, A.J. (1995) Preliminary study of spin-up processes in land surface models with the first stage data of project for Intercomparison of land surface parameterization schemes phase 1(a). *Journal of Geophysical Research: Atmospheres*, 100, 16553–16578. Available from: https://doi.org/10.1029/95JD01076

Zhao, L., Oleson, K., Bou-Zeid, E., Krayenhoff, E.S., Bray, A., Zhu, Q. et al. (2021) Global multi-model projections of local urban climates. *Nature Climate Change*, 11, 152–157. Available from: https://doi.org/10.1038/s41558-020-00958-8

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: Lipson, M.J., Grimmond, S., Best, M., Abramowitz, G., Coutts, A., Tapper, N. *et al*. (2024) Evaluation of 30 urban land surface models in the Urban-PLUMBER project: Phase 1 results. *Quarterly Journal of the Royal Meteorological Society*, 150(758), 126–169. Available from: https://doi.org/10.1002/qj.4589

APPENDIX A1

Error metric definitions

TABLE A1 Error statistics and metrics used to evaluate models

Metric/ statistic	Abbreviation/symbol	Formula	Source
Statistical measures			
Mean absolute error	MAE	$\sum \frac{\lvert M_i - O_i \rvert}{n}$	–
Mean bias error	MBE	$\sum \frac{M_i - O_i}{n}$	–
Pearson correlation coefficient	r	$\frac{\sum \left(M_i - \overline{M}\right)\left(O_i - \overline{O}\right)}{\sqrt{\sum \left(M_i - \overline{M}\right)^2}\sqrt{\sum \left(O_i - \overline{O}\right)^2}}$	–
Standard deviation	σ_X	$\sqrt{\frac{\sum \left(X_i - \overline{X}\right)^2}{n-1}}$	–
Normalised standard deviation	$\tilde{\sigma}$	$\frac{\sigma_M}{\sigma_O}$	(Taylor, 2001)
Skewness	μ_X	$\frac{1}{n}\sum \left(\frac{X_i - \overline{X}}{\sigma_X}\right)^3$	(Best *et al.*, 2015)
Kurtosis	K_X	$\frac{1}{n}\sum \left(\frac{X_i - \overline{X}}{\sigma_X}\right)^4 - 3$	(Best *et al.*, 2015)
Perkins skill score	PSS	$\sum_1^{100} \min\left(bin_{M,k}, \mathrm{bin}_{O,k}\right)$	(Perkins *et al.*, 2007)
Centred and normalisedroot-mean-square error	cRMSE	$\sqrt{1 + \tilde{\sigma}^2 - 2\tilde{\sigma} \cdot r}$	(Taylor, 2001)
Common group metrics			
Mean absolute error metric	m_{MAE}	MAE	–
Mean bias error metric	m_{MBE}	|MBE|	(Best *et al.*, 2015)
Normalised standard deviation metric	m_{SD}	$\lvert 1 - \tilde{\sigma} \rvert$	(Best *et al.*, 2015)
Correlation coefficient metric	m_R	$1 - r$	(Haughton *et al.*, 2017)
Extremes group metrics			
Absolute error at the fifth percentile	m_5	$\lvert M_5 - O_5 \rvert$	(Best *et al.*, 2015)
Absolute error at the 95th percentile	m_{95}	$\lvert M_{95} - O_{95} \rvert$	(Best *et al.*, 2015)
Distribution group metrics			
Skewness metric	m_{skewness}	$\left\lvert 1 - \frac{\mu_M}{\mu_O} \right\rvert$	(Haughton *et al.*, 2017)
Kurtosis metric	m_{kurtosis}	$\left\lvert 1 - \frac{K_M}{K_O} \right\rvert$	(Haughton *et al.*, 2017)
Overlap metric	m_{overlap}	1 − PSS	(Haughton *et al.*, 2017)

Note: Metrics (m) used in group scores are normalised to be positive with 0 a perfect score. M represents modelled values, and O the observed values. An overbar (e.g., $\overline{O}$) indicates the mean of n samples. n varies with observational availability. X_5 is value of X at the fifth percentile of its distribution. Note that for the purposes of scoring models relative to benchmarks, the mean absolute error metric gives equivalent results to the normalised mean error metric used in PLUMBER (Best *et al.*, 2015).

Benchmark observational data

TABLE A2 Site locations of tower observational data used to generate the empirical benchmarks for this study

Sitename	City	Country	Observed period	Latitude	Longitude	References
AU-Preston	Melbourne	Australia	Aug 2003–Nov 2004	−37.7306	145.0145	(Coutts *et al.*, 2007a, 2007b)
AU-SurreyHills	Melbourne	Australia	Feb 2004–Jul 2004	−37.8265	145.099	(Coutts *et al.*, 2007a, 2007b)
CA-sunset	Vancouver	Canada	Jan 2012–Dec 2016	49.2261	−123.078	(Christen *et al.*, 2011; Crawford and Christen, 2015)
FI-Kumpula	Helsinki	Finland	Dec 2010–Dec 2013	60.2028	24.9611	(Karsisto *et al.*, 2016)
FI-Torni	Helsinki	Finland	Dec 2010–Dec 2013	60.1678	24.9387	(Nordbo *et al.*, 2013; Järvi *et al.*, 2018)
FR-Capitole	Toulouse	France	Feb 2004–Mar 2005	43.6035	1.4454	(Masson *et al.*, 2008; Goret *et al.*, 2019)
GR-HECKOR	Heraklion	Greece	Jun 2019–Jun 2020	35.3361	25.1328	(Stagakis *et al.*, 2019)
JP-Yoyogi	Tokyo	Japan	Mar 2016–Mar 2020	35.6645	139.6845	(Hirano *et al.*, 2015; Ishidoya *et al.*, 2020)
KR-Jungnang	Seoul	South Korea	Jan 2017–Apr 2019	37.5907	127.0794	(Hong *et al.*, 2020, 2023)
KR-Ochang	Ochang	South Korea	Jun 2015–Jul 2017	36.7197	127.4344	(Hong *et al.*, 2019, 2020)
MX-Escandon	Mexico City	Mexico	Jun 2011–Sep 2012	19.4042	−99.1761	(Velasco *et al.*, 2011, 2014)
NL-Amsterdam	Amsterdam	Netherlands	Jan 2019–Oct 2020	52.3665	4.8929	(Steeneveld *et al.*, 2020)
PL-Lipowa	Łódź	Poland	Jan 2008–Dec 2012	51.7625	19.4453	(Pawlak *et al.*, 2011; Fortuniak *et al.*, 2013)
PL-Narutowicza	Łódź	Poland	Jan 2008–Dec 2012	51.7733	19.4811	(Fortuniak *et al.*, 2013, 2006)
SG-TelokKurau06	Singapore	Singapore	Apr 2006–Mar 2007	1.3143	103.9112	(Roth *et al.*, 2017)
UK-KingsCollege	London	UK	Apr 2012–Jan 2014	51.5118	−0.1167	(Kotthaus and Grimmond, 2014a, 2014b; Bjorkegren *et al.*, 2015)
UK-Swindon	Swindon	UK	May 2011–Apr 2013	51.5846	−1.7981	(Ward *et al.*, 2013)
US-Baltimore	Baltimore	USA	Jan 2002–Jan 2007	39.4128	−76.5215	(Crawford *et al.*, 2011)
US-Minneapolis	Minneapolis	USA	Jun 2006–May 2009	44.9984	−93.1884	(Peters *et al.*, 2011; Menzer and McFadden, 2017)
US-WestPhoenix	Phoenix	USA	Dec 2011–Jan 2013	44.9984	−93.1884	(Chow *et al.*, 2014; Chow, 2017)

Note: For the out-of-sample benchmarks, data from the AU-Preston site are not used to train the empirical models. For the in-sample benchmarks, only AU-Preston data are used to train the model. Tower data are openly available (Lipson *et al.*, 2022a, 2022c).

Participating model descriptions

For each model (Figures A1–A30) characteristics (Figure 1) are summarised. Table 2 gives the version of the model used in this paper. Diagrams are indicative only. Individual models have attributes not represented in diagrams.

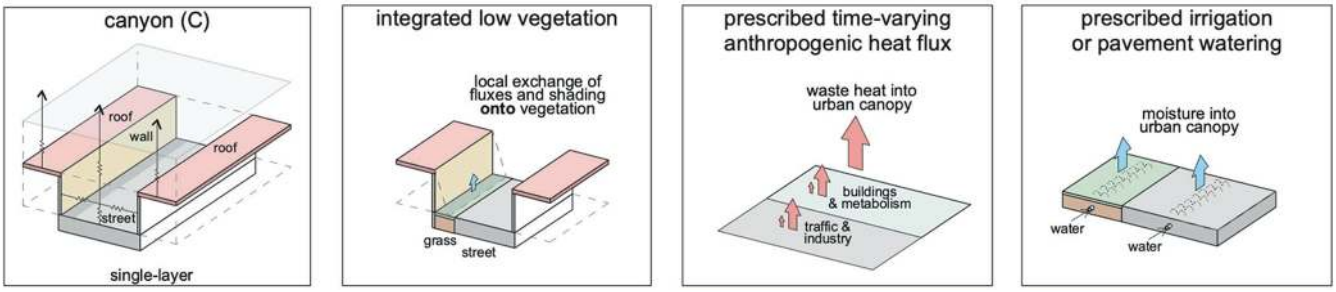

FIGURE A1 ASLUMv2.0 (Arizona State University Single-Layer Urban Canopy Model) (Wang *et al.*, 2021, 2013) is a single-layer canyon model that analytically resolves surface temperatures and conductive heat fluxes based on Green's function (Wang *et al.*, 2011b), and explicitly resolves subfacet heterogeneity and urban vegetation. ASLUMv2 incorporates detailed hydrology, multilayer soil/ground and roof vegetation (via a multilayer green roof model), and enables simulations of irrigation, anthropogenic heat, and urban oasis effect (Yang *et al.*, 2015).

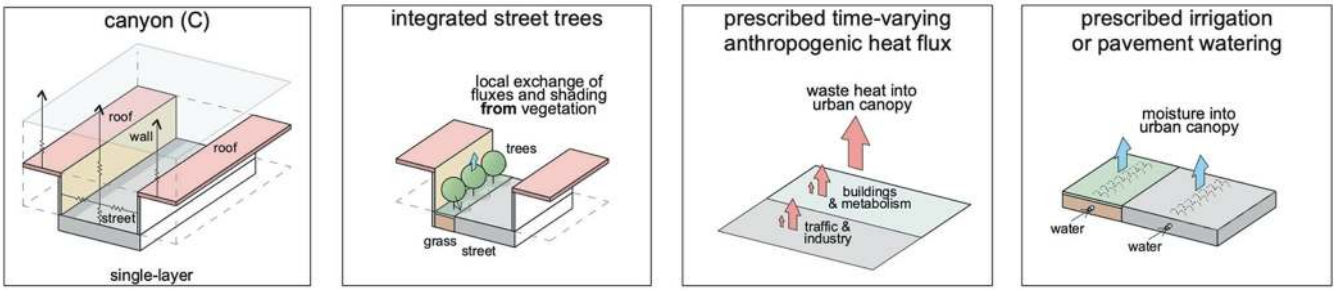

FIGURE A2 ASLUMv3.1 (Wang *et al.*, 2021, 2013) is the same as ASLUMv2.0 (Figure A1) plus additionally represents urban trees with radiative exchange between trees and street canyons (Wang, 2014), shading, canopy transmittance, evapotranspiration, and root water uptake (Wang *et al.*, 2021).

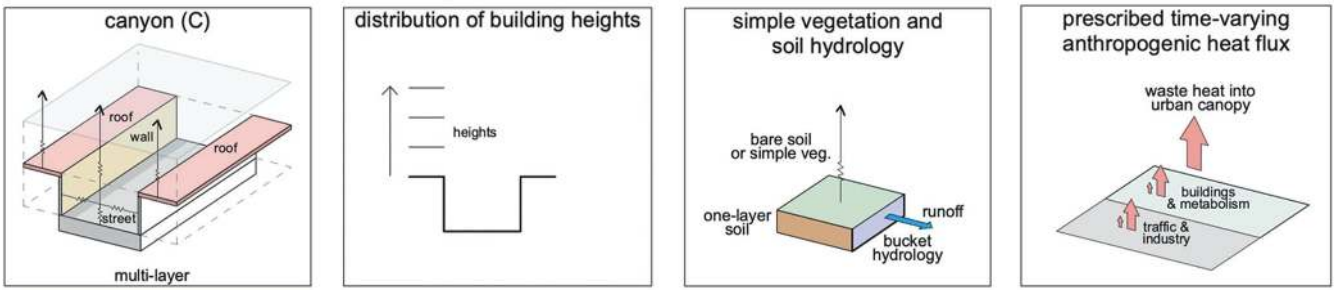

FIGURE A3 BEPCOL is a multilayer canyon model based on the building effect parameterization (BEP; Martilli *et al.*, 2002) with parameterizations for drag coefficient and the length scales used for turbulent transport and turbulence dissipation (Simón-Moral *et al.*, 2017). BEPCOL does not explicitly consider vegetation and the non-urban fraction is computed by the bare soil model from RAMS (Tremback and Kessler, 1985).

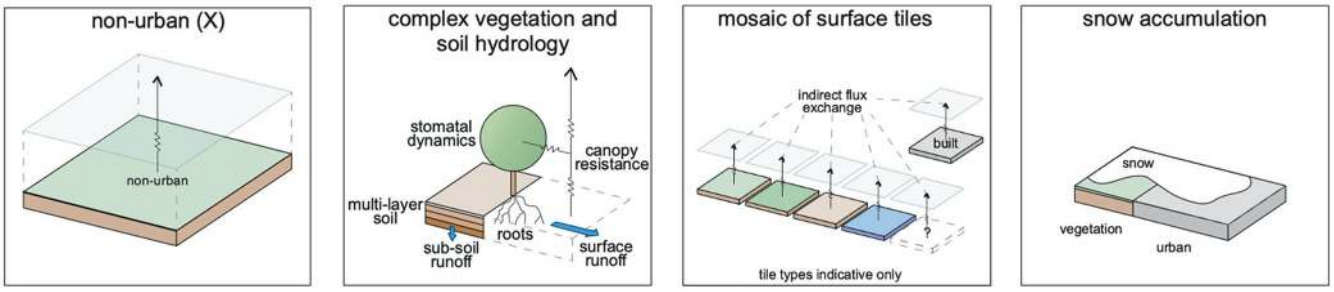

FIGURE A4 CABLE (Community Atmosphere–Biosphere Land Exchange model) (Kowalczyk *et al.*, 2006; Wang *et al.*, 2011a) is used in regional and global climate models including ACCESS (Kowalczyk *et al.*, 2013). CABLE has a one-layer, two-leaf canopy vegetation scheme with up to five tiles (vegetation types, bare soil and ice) but no urban tile. In this project CABLE uses four soil layers, up to three snow layers, and models impervious urban surfaces as bare soil.

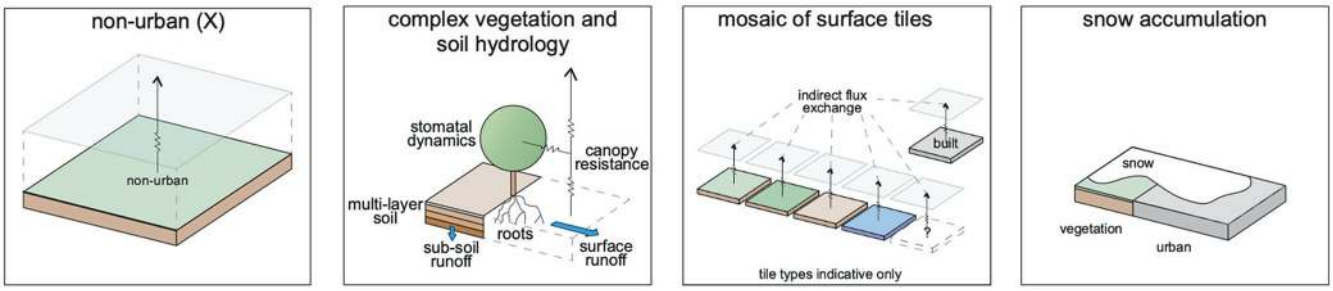

FIGURE A5 CHTESSEL (Carbon-Hydrology Tiled ECMWF Forecasts Scheme for Surface Exchanges over Land) is the land surface model used in the Integrated Forecast System (IFS). This is used by ECMWF for weather forecast and to create reanalysis products (Balsamo *et al.*, 2009; Boussetta *et al.*, 2013; ECMWF, 2020). CHTESSEL can tile up to six non-urban surfaces (high and low vegetation, bare soil, intercepted canopy water, shaded and sunlit snow). It has four soil and one snow layer. Tile fractions in this project are based on global surface cover databases as used in IFS products, i.e., without urban surfaces.

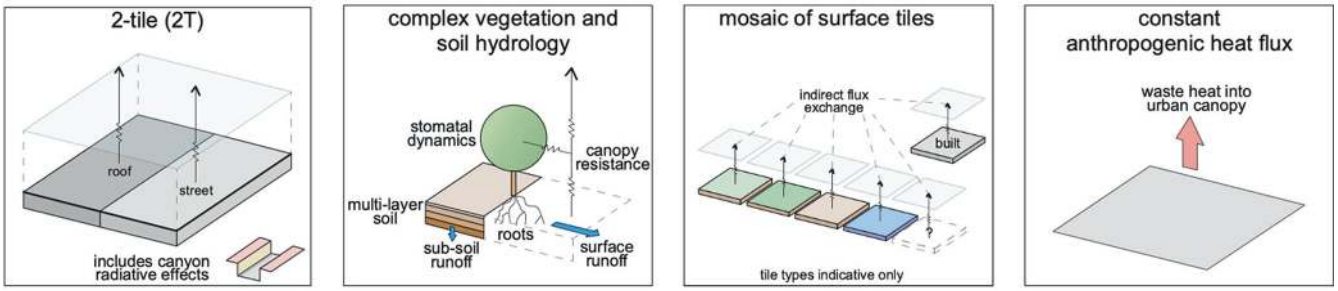

FIGURE A6 CHTESSEL_U (Urbanised Carbon-Hydrology Tiled ECMWF Forecasts Scheme for Surface Exchanges over Land) a two-tile (roof, canyon) urban scheme (McNorton *et al.*, 2021) to CHTESSEL (Figure A5). It follows MORUSES' (Figure A13) infinite canyon assumptions for radiative effects. The urban surfaces (cf. CHTESSEL) have increased runoff and reduced soil infiltration.

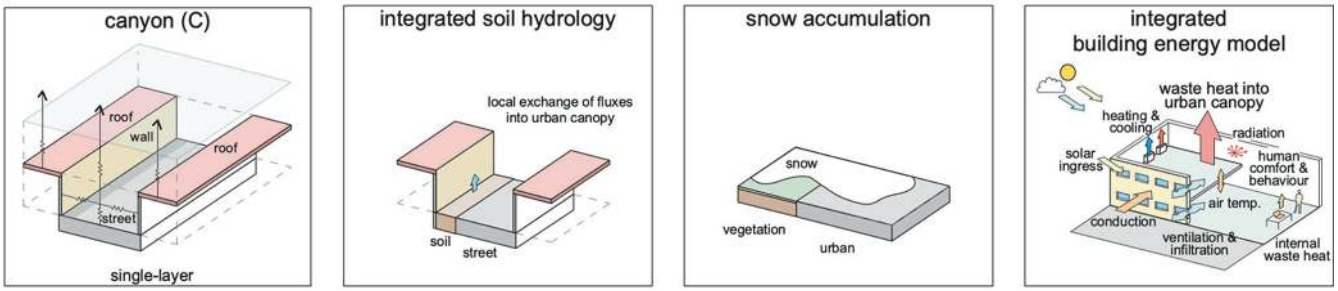

FIGURE A7 The CLMU (Community Land Model Urban) is a single-layer urban canopy model that consists of roofs, walls (sunlit and shaded) and impervious and pervious canyon floor (Oleson *et al.*, 2010; Oleson and Feddema, 2020). It features a simple building energy model to explicitly represent space heating and air conditioning of building interiors and a pervious canyon floor to approximate evaporation from vegetated surfaces within the urban canyon. Radiation parameterizations account for trapping of short-wave and long-wave radiation inside the urban canyon. Roof and canyon floor hydrologic processes including snow accumulation and melt, liquid water ponding, and runoff are simulated. CLMU is embedded within a global climate model, the Community Earth System Model (CESM), incorporating three urban tiles or land units, categorized by density of development, within each model grid cell. Urban extent and thermal, radiative, and morphological properties are prescribed from the global dataset of Jackson *et al.* (2010) as modified by Oleson and Feddema (2020).

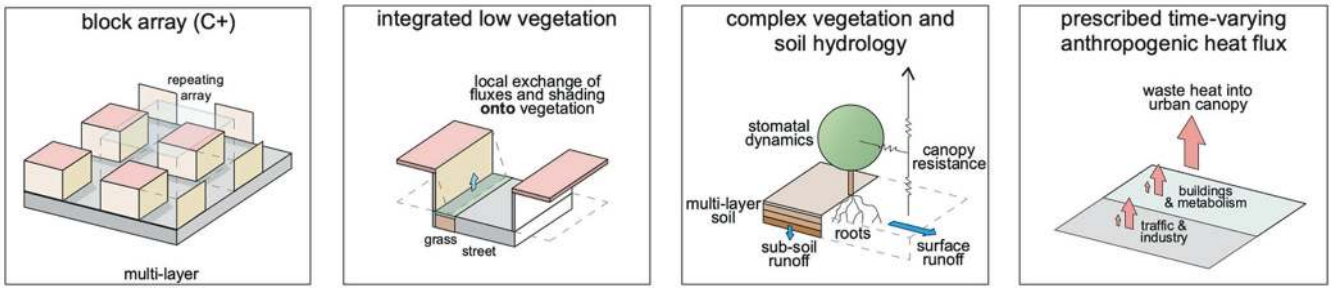

FIGURE A8 CM is a multilayer urban canopy model with roofs (impervious roof and vegetation), walls (impervious wall and vegetation) and roads (impervious road and vegetation) (Kondo and Liu, 1998; Kondo *et al.*, 2005). These impervious tiles consider water content and ponding for the present comparison. CM considers an urban block in which buildings stand on a lattice array. This horizontal arrangement of buildings is defined using the average length of the building and distance between buildings. CM accounts for building drag, anthropogenic heat release (prescribed), and three-dimensional radiation interactions and distribution of the height of buildings. A new parameterization for mixing length is introduced (Kondo *et al.*, 2015).

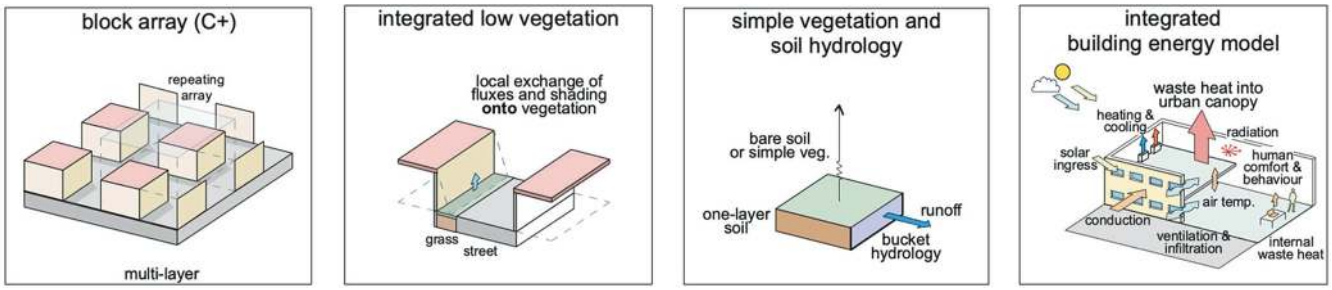

FIGURE A9 CM-BEM is CM (Figure A8) coupled to a building energy model (BEM) (Kikegawa *et al.*, 2003, 2006). It is embedded within WRF (WRF-CM-BEM: Kikegawa *et al.*, 2014). It has three urban categories: office, and two residential types. The BEM, box-type heat budget model, simulates heating, ventilation, and air conditioning (HVAC) system energy consumption and the resulting anthropogenic heat release including sensible and latent heat components. BEM can consider whether the outdoor units are air-cooled or water-cooled. CM-BEM can integrate social big data such as real-time population and estimate the impacts of human behaviour changes on urban temperature and energy consumption (Takane *et al.*, 2022).

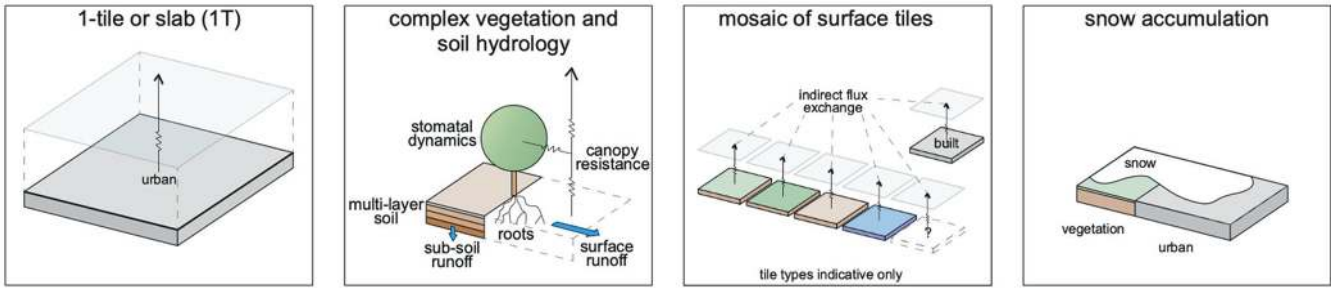

FIGURE A10 JULES_1T is a one-tile urban scheme (Best, 2005) within the Joint UK Land Environment Simulator (JULES) (Best *et al.*, 2011). JULES is a community land surface model forming part of the UK Met Office's Unified Model (UM). JULES has nine non-urban surface tiles (five vegetation types, inland water, bare soil and ice). Tiles can be covered with up to three snow layers. Soil hydro-thermodynamics are modelled through four soil layers, with the top layer coupled to the urban tile through long-wave radiation only.

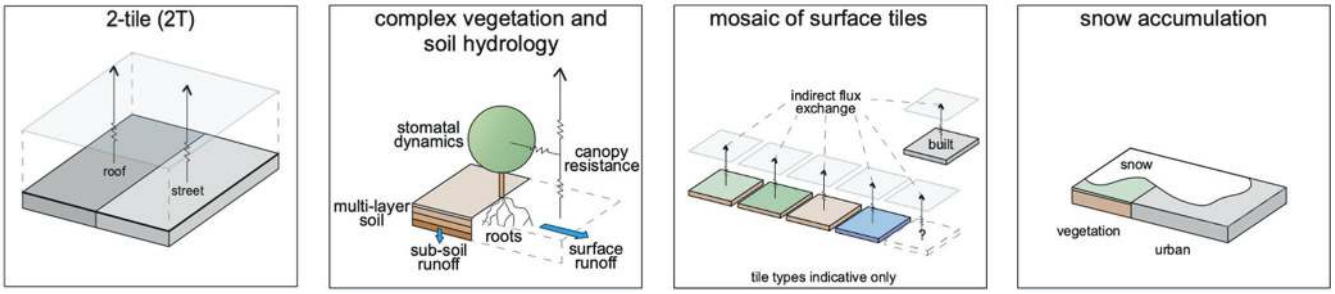

FIGURE A11 JULES_2T is a two-urban tile (roof, canyon) (Best *et al.*, 2006) version of JULES_1T (Figure A10).

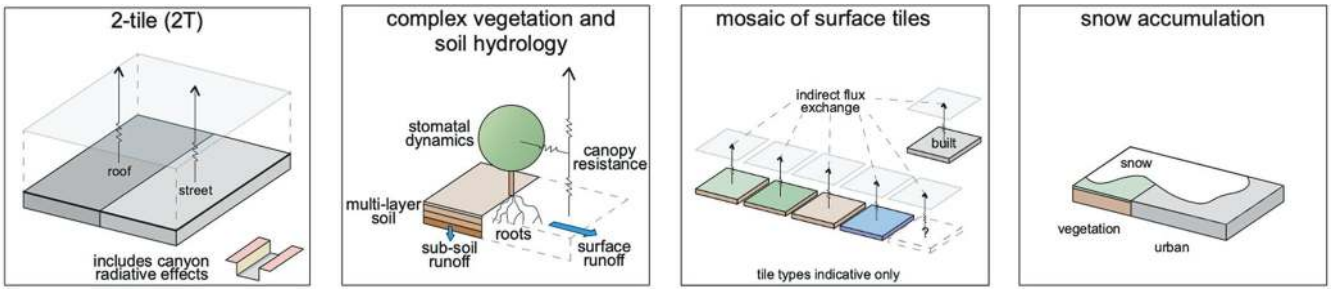

FIGURE A12 JULES_MORUSES, has the same two urban tiles (roof, canyon) as JULES_2T (Figure A11) but uses MORUSES (Porson *et al.*, 2010). MORUSES has only two facets (roof and canyon) but includes a parameterisation for canyon radiative effects which updates canyon bulk values such as albedo, emissivity, heat capacity and aerodynamic resistance. This provides benefits of canyon schemes with a computational efficiency gain important for use in operational numerical weather prediction.

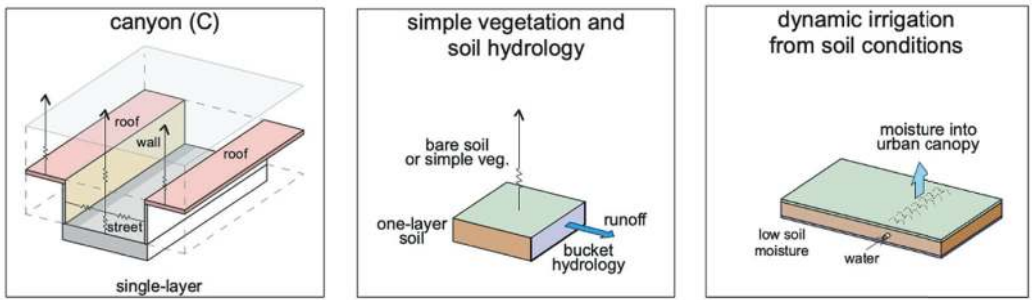

FIGURE A13 K-UCMv1 (Klimaat Urban Canopy Model v1) solves the local surface energy balance based on land cover input and regional meteorological forcing. To calculate conduction, the urban facets (ground, roof, walls) have 10 layers. Effects of low vegetation are accounted for in the surface energy balance without shadowing effects. Vegetation is perfectly watered and non-vegetation surfaces store no water.

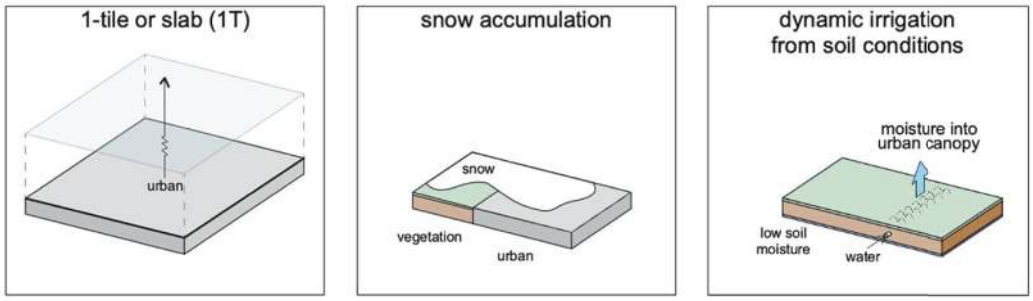

FIGURE A14 Lodz-SUEB (Łódź SUrface Energy Balance model) (Fortuniak, 2003) is a bulk scheme that aggregates urban and natural surfaces parameters based on surface fraction. The ground heat flux assumes a 12-layer slab. This is a single snow layer and any surface liquid water or snow, if present, assumed to completely cover the slab. Moisture content on the slab is constrained between site-specific limits, with excess leading to runoff and supply from deeper layers during dry periods.

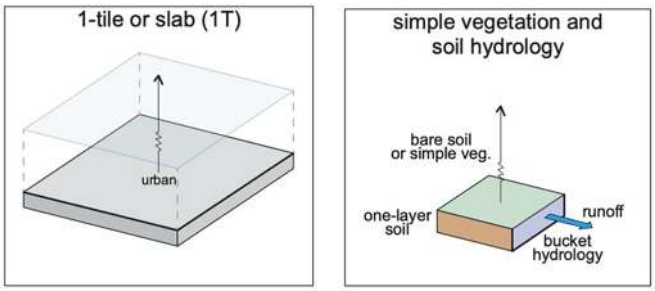

FIGURE A15 Manabe_1T uses a one-tile urban scheme (Best, 2005) with a simple Manabe bucket model for non-urban fractions (Manabe, 1969). A Manabe bucket model has no heat conduction into the soil and accumulates precipitation until it freely evaporates or exceeds storage capacity as runoff (Pitman, 2003). We use this 'slab and bucket' scheme as a physically-based benchmark.

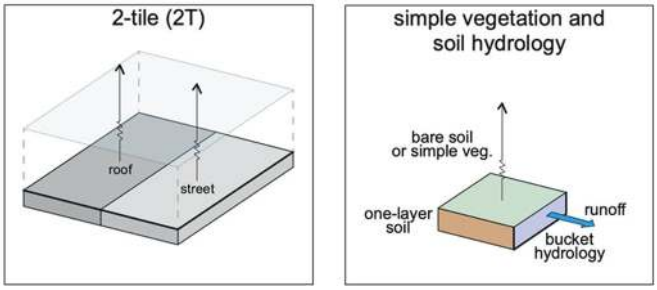

FIGURE A16 Manabe_2T only differs from Manabe_1T (Figure A15) in that it uses a two-tile urban scheme (roof, canyon) (Best *et al.*, 2006).

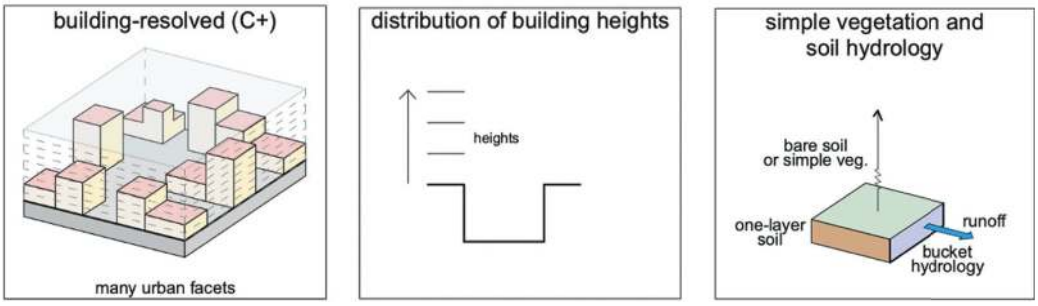

FIGURE A17 MUSE (Microscale Urban Surface Energy) (Lee and Lee, 2020) is a building-resolving microscale urban surface model for real urban meteorological and environmental applications. It represents urban buildings on a three-dimensional Cartesian grid and solves urban physical processes of short-wave and long-wave radiative transfer, turbulent exchanges of momentum and heat, and conductive heat transfer into urban subsurfaces. The effect of urban vegetation is parameterized based on a simple Bowen ratio method in calculating the radiative and turbulent sensible/latent heat fluxes.

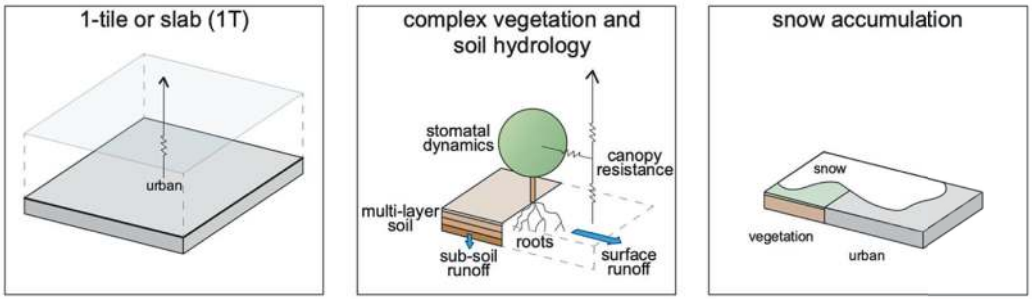

FIGURE A18 NOAH-SLAB uses a bulk one-tile urban scheme (Liu *et al.*, 2006) with the Noah land surface model (Noah-LSM) (Chen and Dudhia, 2001). Separate urban and non-urban energy and water balances are simulated, then weighted by surface fraction. Although Noah-LSM has up to 27 land-use tiles (including urban), here the urban and one dominant non-urban land-uses are used. For the non-urban surface, parameters (e.g., albedo, roughness length) are set to urban values provided, allowing evaporation from vegetation for an otherwise urban surface.

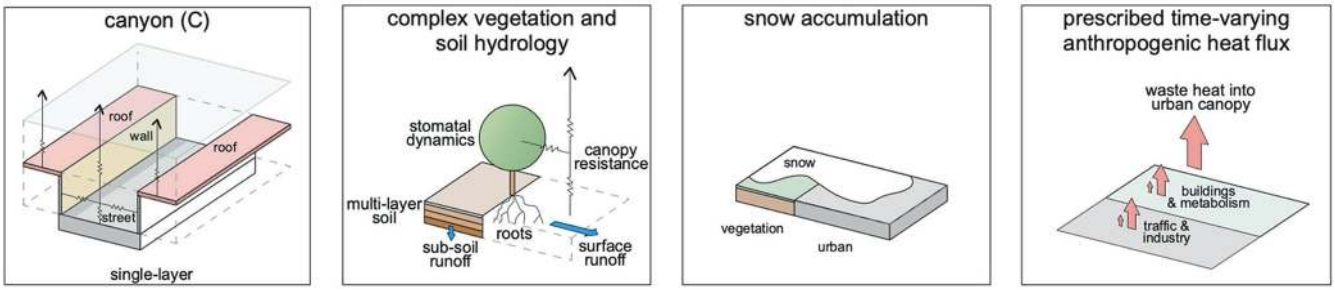

FIGURE A19 NOAH-SLUCM uses Noah-LSM as in NOAH-SLAB (Figure A18) but uses the Single-Layer Urban Canopy Model (Kusaka *et al.*, 2001; Chen *et al.*, 2011) rather than the urban slab scheme. SLUCM separates the urban tile into three facets (roof, road, wall) using a two-dimensional canyon approach but without street orientation or varying building heights. A diurnally varying anthropogenic heat flux is prescribed.

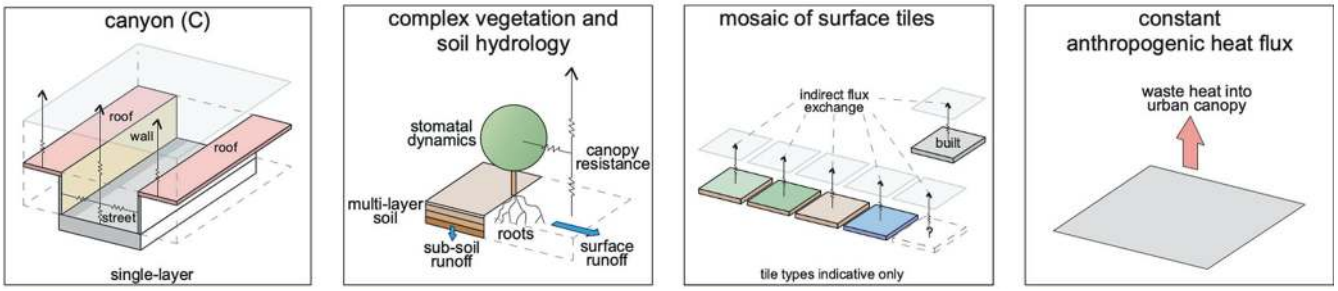

FIGURE A20 SNUUCM (Seoul National University Urban Canopy Model), a single-layer model, parameterises short-wave and long-wave radiation absorption and reflection, energy and moisture turbulent exchanges between surfaces and adjacent air, and conductive heat transfer through sublayers (Ryu *et al.*, 2011). It calculates canyon wind speed using regression equations based on CFD model simulations. Here the non-urban area fluxes are simulated by the Noah land surface model v3.4.1 (Chen and Dudhia, 2001).

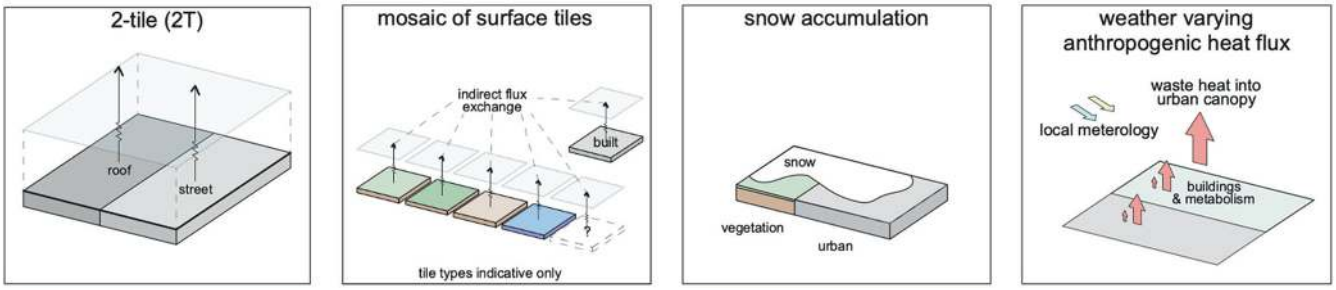

FIGURE A21 SUEWS (Surface Urban Energy and Water Balance Scheme) (Järvi *et al.*, 2011; Ward *et al.*, 2016) has two impervious (buildings, paved areas) and five pervious (evergreen trees and shrubs, deciduous trees and shrubs, grass, bare soil and water) surface types, underneath which is a single vertical layer for soil model with lateral flow between surfaces (except water tile). Storage heat fluxes can be calculated using empirical relations with net all wave radiation (Grimmond *et al.*, 1991) while the latent heat flux is calculated as the integrated resistance network of all the surfaces. There is one snow layer but with clearance activities between surfaces. Anthropogenic heat emissions, irrigation-related fluxes and snow-clearing are either modelled with empirical relations or prescribed with observed values. It has a dynamic leaf area index model to allow phenology to change through the year and between years.

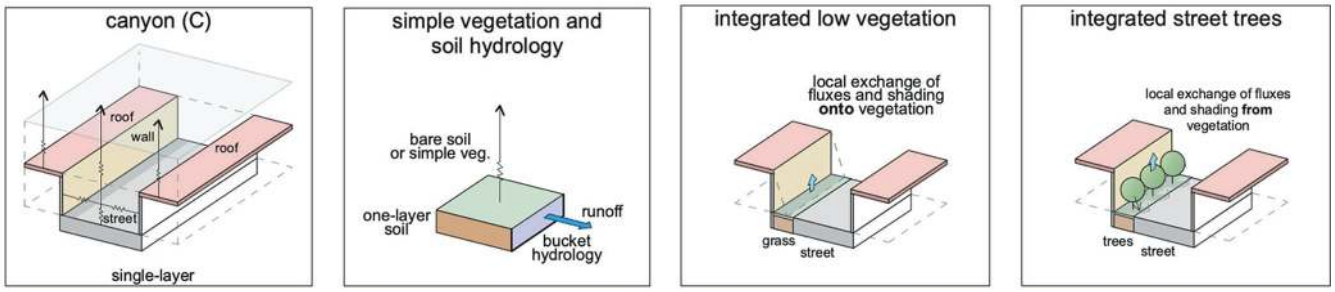

FIGURE A22 TARGET (The Air temperature Response to Green/blue-infrastructure Evaluation Tool) (Broadbent *et al.*, 2019) models the canyon-to-block scale street-level air temperature impacts of green/blue infrastructure. Grid points are represented as idealised urban canyons using width/height to define the geometry and an aggregate of land cover surface types (concrete, asphalt, grass, irrigated grass, vegetation, and water). TARGET is designed to predict street-level conditions, not bulk surface–atmosphere fluxes.

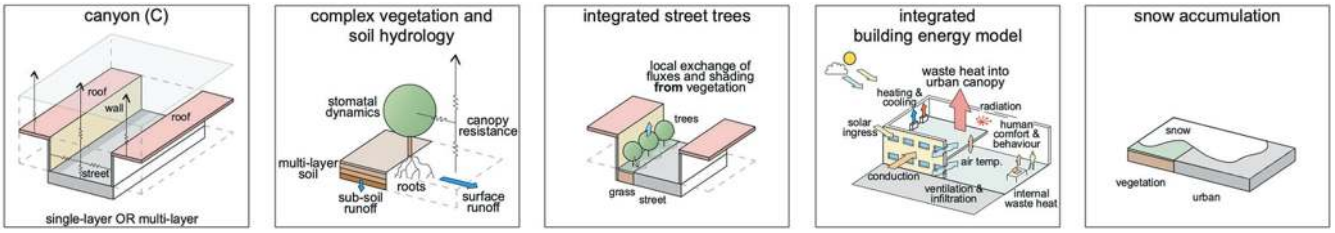

FIGURE A23 TEB-CNRM uses the multilayer version (Hamdi and Masson, 2008; Schoetter *et al.*, 2017) of the urban canopy model Town Energy Balance (Masson, 2000) and part of Météo-France's SURFEX Land Surface Model (Masson *et al.*, 2013; Le Moigne *et al.*, 2018). Here TEB buildings (roofs, walls), roads and urban vegetation on the ground (grass, shrubs) (Lemonsu *et al.*, 2012) influence each other directly. Wind effects are averaged assuming all street orientations exist. Water and snow can be present on roofs, roads and urban vegetation. Street trees are treated as an elevated tree-foliage stratum that partially covers the ground and shadows walls and ground surfaces (Redon *et al.*, 2017). Soil hydrology is resolved in three soil compartments below vegetation, roads and buildings (Stavropulos-Laffaille *et al.*, 2018, 2021). A Building Energy Model (BEM) is included with human behaviour (Bueno *et al.*, 2012; Schoetter *et al.*, 2017).

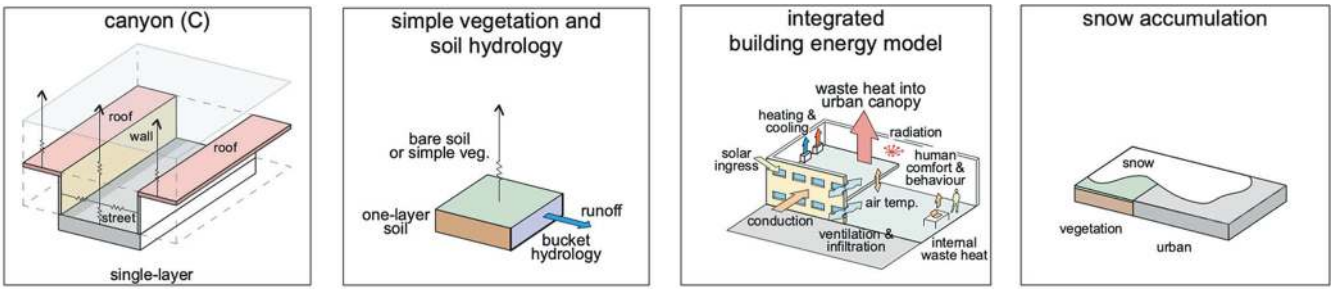

FIGURE A24 TEB-READING uses the offline single-layer Town Energy Balance model (Masson, 2000) software (Meyer *et al.*, 2020a) version 4.1.0 (Masson *et al.*, 2021). TEB 4.1.0 is similar to the single layer TEB used in SURFEX version 8.1 (Le Moigne *et al.*, 2018) but with a simple vegetation scheme and time-constant Bowen ratio, albedo, roughness, soil temperature and water availability (Meyer *et al.*, 2020a, 2020b). This simplified vegetation scheme neglects heat conduction and assumes neutral conditions for friction velocity. The Building Energy Model by Bueno *et al.* (2012) uses MinimalDX (Meyer and Raustad, 2020) to improve air conditioners' modelling.

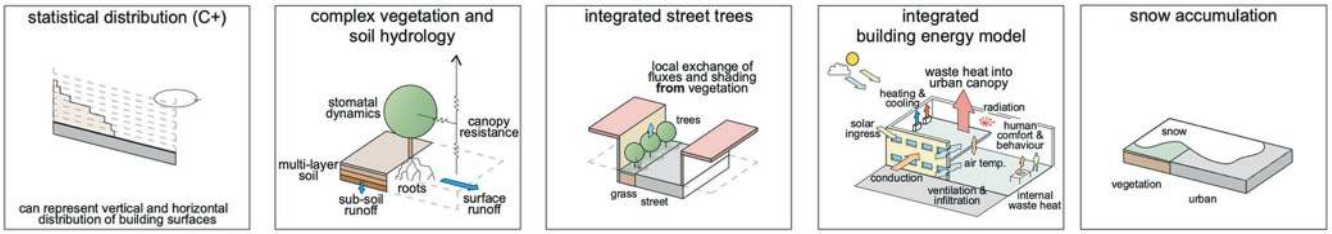

FIGURE A25 TEB-SPARTCS is builds on TEB-CNRM (Figure A23) by incorporating the SPARTACUS-Urban radiative transfer within the urban canopy using a discrete ordinate method (Hogan, 2019a, 2019b), which assumes an exponential distribution of wall-to-wall distances and allows varying building heights (Hogan, 2019a). In this project the buildings all have the same height and tree height is limited to building height.

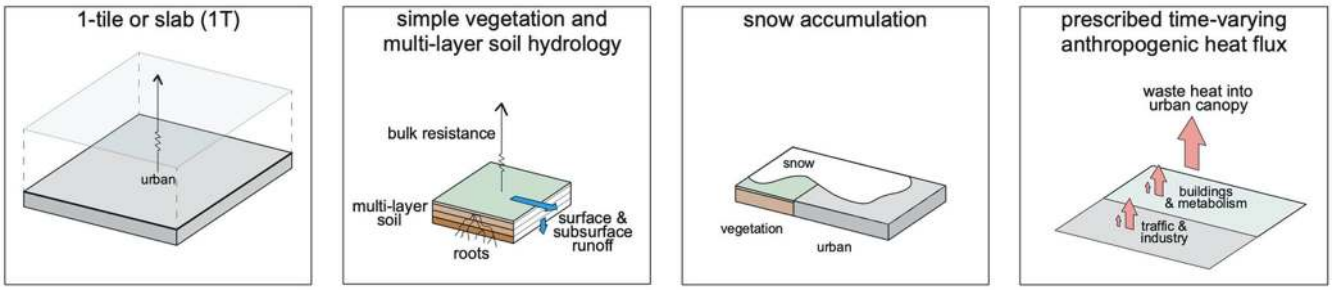

FIGURE A26 TERRA_4.11 uses the bulk urban scheme TERRA_URB (Wouters *et al.*, 2015, 2016) with TERRA-ML (Schulz *et al.*, 2016) for non-urban surfaces that are characterized by eight soil layers and a one-layer snow scheme. As the latter does not account for urban features (e.g., urban pollution, snow removal or a change in effective albedo due to snow-free walls, roads), the urban fraction is considered as completely snow-covered in the presence of snowfall. TERRA_URB uses the Semi-empirical URban canopy dependencY algorithm (SURY) to condense the three-dimensional urban canopy information to a limited number of bulk properties (Wouters *et al.*, 2016; Varentsov *et al.*, 2020). Aggregated diurnal and seasonal anthropogenic heat fluxes (traffic, industry, etc. combined) are prescribed by equations proposed in Flanner (2009). The TERRA version used here is a standalone version, which differs from the official TERRA version embedded in the recent COSMO(-CLM) version (Rockel *et al.*, 2008; Garbero *et al.*, 2021) (Rockel *et al.*, 2008; Garbero *et al.*, 2021). Where possible, features from the online version are approximated using the same underlying data sources. For more information on this submission see https://github.com/matthiasdemuzere/urban-plumber-terra-pub.

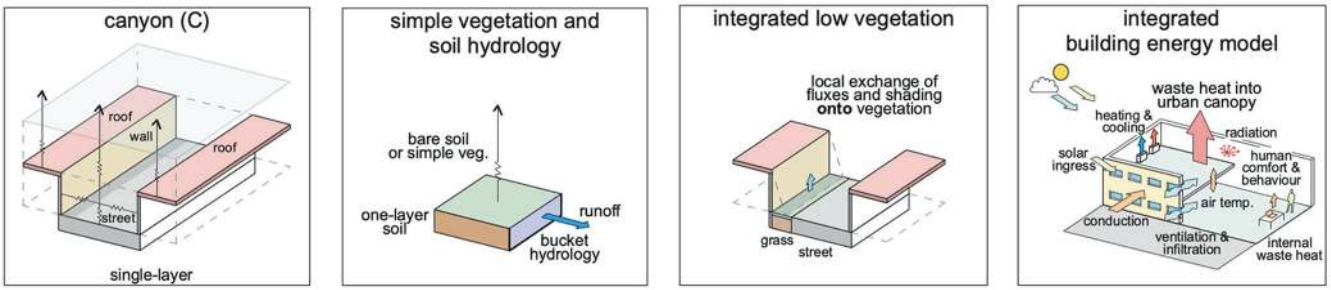

FIGURE A27 UCLEM (Urban Climate and Energy Model), with integrated street vegetation and building energy/waste heat (Lipson et al., 2018; Thatcher and Hurley, 2012), is used in the stretched grid global climate model CCAM (McGregor and Dix, 2008). The four urban facets (roof, road, two walls) have four layers/ five nodes for heat conduction (Lipson *et al.*, 2017), with single-layer snow on a fraction of roof and road surfaces. Low (grass and shrub) canyon and roof vegetation use a reduced set of prognostic variables with a simple bucket hydrology. Irrigation is assumed to occur when soil moisture approaches wilting point.

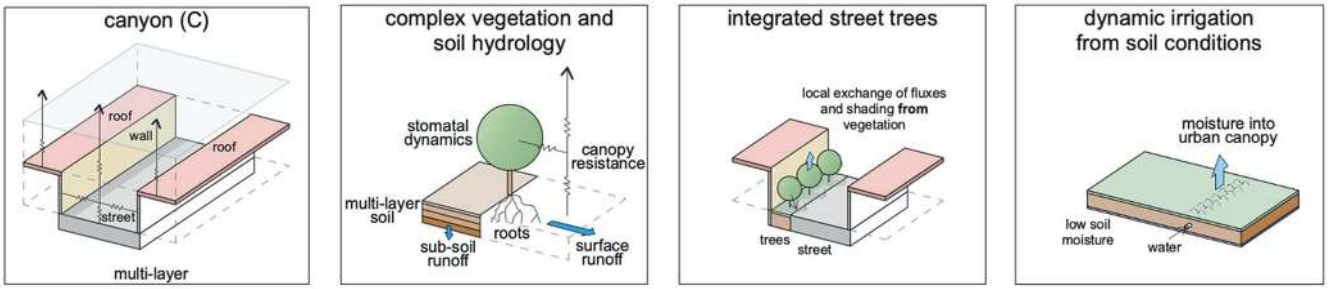

FIGURE A28 UT&C: (Urban Tethys-Chloris) (Meili *et al.*, 2020) combines an urban canyon approach with ecohydrological principles of Tethys-Chloris (Fatichi *et al.*, 2012). Vegetation can occur on roofs and within the canyon (i.e., ground vegetation and/or street trees). Separate soil columns occur below impervious, bare and vegetated ground facets. Transpiration is modelled as a function of plant photosynthetic activity and environmental conditions (Meili *et al.*, 2021). Irrigation can be prescribed at the soil surface or through preserving soil moisture in deep soil layers. Wall facets are split into upper and lower parts to partition their contribution to near surface heat fluxes. Snow and water bodies are currently not modelled in UT&C v1.0.

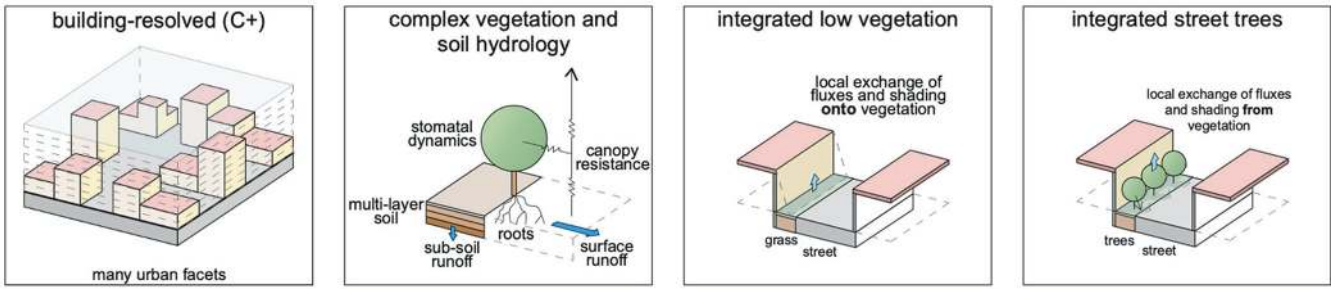

FIGURE A29 VTUF-3D (Vegetated Temperatures of Urban Facets in 3D) (Nice *et al.*, 2018) resolves energy transfers in three dimensions by combining TUF-3D (Krayenhoff and Voogt, 2007) with the MAESPA tree model (Duursma and Medlyn, 2012). The vegetation shading and physiological processes are directly integrated with building and urban effects, allowing the role vegetation and water to be assessed in human thermal comfort in urban areas.

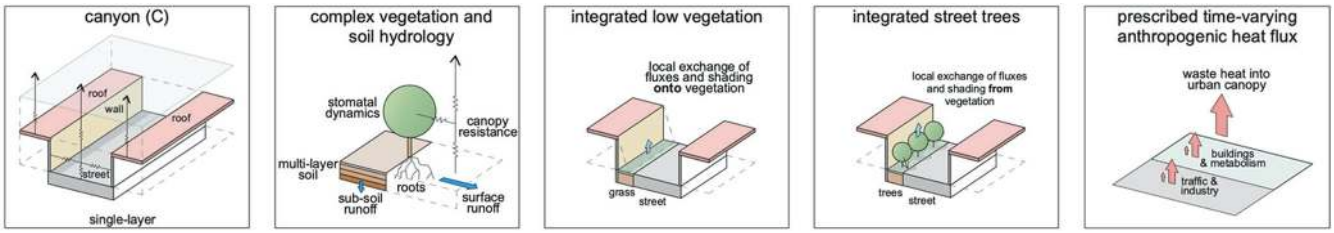

FIGURE A30 VUCM (Vegetated Urban Canopy Model) (Lee and Park, 2008; Lee, 2011; Lee *et al.*, 2016) is the first mesoscale urban canopy model that parametrizes radiative/dynamic/thermodynamic/hydrological processes of urban vegetated area (tree, grass, soil) interactively with urban artificial surfaces (roof, wall, road), which has been developed based on an integrated framework of a two-dimensional single canyon and a new single tree canopy concept.