Lipson Mathew Jack (Orcid ID: 0000-0001-5322-1796)
Grimmond Sue (Orcid ID: 0000-0002-3166-9415)
Baik Jong-Jin (Orcid ID: 0000-0003-3709-0532)
Blunn Lewis (Orcid ID: 0000-0002-3207-5002)
Hendry Margaret A (Orcid ID: 0000-0003-3941-7543)
Meili Naika (Orcid ID: 0000-0001-6283-2134)
Meyer David (Orcid ID: 0000-0002-7071-7547)
Nice Kerry A (Orcid ID: 0000-0001-6102-1292)
Oleson Keith W (Orcid ID: 0000-0002-0057-9900)
Roth Michael (Orcid ID: 0000-0001-6399-3693)
Simon-Moral Andres (Orcid ID: 0000-0002-2662-9750)
Steeneveld Gert-Jan (Orcid ID: 0000-0002-5922-8179)
Sun Ting (Orcid ID: 0000-0002-2486-6146)
Wang Chenghao (Orcid ID: 0000-0001-8846-4130)
Wang Zhihua (Orcid ID: 0000-0001-9155-8605)

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

Mathew J. Lipson[1], Sue Grimmond[2], Martin Best[3], Gab Abramowitz[4], Andrew Coutts[5], 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[18], Yukihiro Kikegawa[19], Hiroaki Kondo[20], Doo-Il Lee[21], Sang-Hyun Lee[22], Aude Lemonsu[23], Tiago Machado[24], Gabriele Manoli[25], Alberto Martilli[26], Valéry Masson[27], Joe McNorton[28], Naika Meili[29], David Meyer[30], Kerry A. Nice[31], Keith W. Oleson[32], Seung-Bu Park[33], Michael Roth[34], Robert Schoetter[35], Andrés Simón-Moral[36], Gert-Jan Steeneveld[37], Ting Sun[38], Yuya Takane[39], Marcus Thatcher[40], Aristofanis Tsiringakis[41], Mikhail Varentsov[42], Chenghao Wang[43], Zhi-Hua Wang[44], Andy J. Pitman[45]

Corresponding author: Mathew J. Lipson (mathew.lipson@bom.gov.au)

[1] 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.

[2] Department of Meteorology, University of Reading, Reading, RG6 6ET, United Kingdom

[3] Met Office, Fitzroy Road, Exeter, Devon, EX1 3PB, United Kingdom

[4] Australian Research Council Centre of Excellence for Climate Extremes, Climate Change Research Centre, Level 4, Mathews Building, UNSW Sydney, New South Wales, 2052, Australia, 0000-0002-4205-001X

[5] School of Earth, Atmosphere and Environment, Monash University, Melbourne, 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, N1E 2K1, Ontario, Canada

[9] Met Office, University of Reading, Reading, RG6 6ET, United Kingdom. 0000-0002-3207-5002

[10] European Centre for Medium-Range Weather Forecasts, Reading, RG2 9AX, United Kingdom

[11] Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ, USA. 0000-0002-6137-8109

[12] School of Biological Sciences, University of Bristol, Bristol, BS8 1TQ, United Kingdom 0000-0002-3399-9098

[13] CNRM, Université de Toulouse, Météo-France, CNRS, 42 Avenue Gaspard Coriolis, 31057 CEDEX 1, Toulouse, France

[14] Urban Climatology Group, Department of Geography, Ruhr-University Bochum, Bochum, Germany 0000-0003-3237-4077

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/qj.4589

[15] Department of Civil and Environmental Engineering, National University of Singapore, Singapore, Singapore, 0000-0003-1361-6659

[16] Department of Meteorology and Climatology, Faculty of Geographical Sciences, University of Lodz, Lodz, Poland. 0000-0001-7043-8751

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

[18] Met Office, Fitzroy Road, Exeter, Devon, EX1 3PB, United Kingdom

[19] School of Science and Engineering, Meisei University, Hino City, Tokyo, Japan. 0000-0002-5225-653X

[20] Environmental Management Research Institute, National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Japan; Japan Weather Association, Tokyo, Japan

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

[22] Department of Atmospheric Science, Kongju National University, Gongju, Republic of Korea. 0000-0002-7998-9194

[23] CNRM, Université de Toulouse, Météo-France, CNRS, 42 Avenue Gaspard Coriolis, 31057 CEDEX 1, Toulouse, France

[24] CNRM, Université de Toulouse, Météo-France, CNRS, 42 Avenue Gaspard Coriolis, 31057 CEDEX 1, Toulouse, France

[25] School of Architecture, Civil and Environmental Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, 0000-0002-9245-2877

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

[27] CNRM, Université de Toulouse, Météo-France, CNRS, 42 Avenue Gaspard Coriolis, 31057 CEDEX 1, Toulouse, France

[28] European Centre for Medium-Range Weather Forecasts, Reading, RG2 9AX, United Kingdom

[29] Department of Civil and Environmental Engineering, National University of Singapore, Singapore, Singapore, 0000-0001-6283-2134

[30] Department of Meteorology, University of Reading, Reading, United Kingdom; Department of Civil and Environmental Engineering, Imperial College London, London, United Kingdom. 0000-0002-7071-7547

[31] Transportation, Health and Urban Design Research Lab, Faculty of Architecture, Building and Planning, University of Melbourne, Australia. 0000-0001-6102-1292

[32] Climate and Global Dynamics Laboratory, National Center for Atmospheric Research, Boulder, CO, USA

[33] School of Environmental Engineering, University of Seoul, South Korea

[34] Klimaat Consulting & Innovation Inc., Guelph, N1E 2K1, Ontario, Canada. 0000-0001-6399-3693

[35] CNRM, Université de Toulouse, Météo-France, CNRS, 42 Avenue Gaspard Coriolis, 31057 CEDEX 1, Toulouse, France

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

[37] Wageningen University, Meteorology and Air Quality Section, Droevendaalsesteeg 3, 6708 PB Wageningen, The Netherlands. 0000-0002-5922-8179

[38] Institute for Risk and Disaster Reduction, University College London, London, WC1E 6BT, United Kingdom. 0000-0002-2486-6146

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

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

[41] European Centre for Medium-Range Weather Forecasts, Bonn, 53175, Germany, 0000-0002-6922-5086

[42] Research Computing Center, Lomonosov, Moscow State University, Moscow, Russia; A.M. Obukhov Institute of Atmospheric Physics, Moscow, Russia, 0000-0001-9095-5334

[43] School of Meteorology, University of Oklahoma, Norman, USA; Department of Geography and Environmental Sustainability, University of Oklahoma, Norman, USA. 0000-0001-8846-4130

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

[45] Australian Research Council Centre of Excellence for Climate Extremes, Climate Change Research Centre, Level 4, Mathews Building, UNSW Sydney, New South Wales, 2052, Australia

Abstract:

Accurately predicting weather and climate in cities is critical for safeguarding human health and strengthening urban resilience. Multi-model evaluations can lead to model improvements, however there have been no major intercomparisons of urban-focused land surface models in over a decade. Here, in Phase 1 of the Urban-PLUMBER project, we evaluate 30 land surface models' ability 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 shortwave radiation, sensible and latent heat fluxes, but little or no improvement in longwave 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 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 twenty sites in different urban and regional climate zones.

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 1) bulk surface albedo during the day, 2) trapping of longwave radiation between urban structures at night, and 3) 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 Parameterization Schemes). Since the 1990's, 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 model comparison, 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 shortwave 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 twenty sites. The PLUMBER project authors 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 1), 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 1) evaluate land surface model performance in an urban setting using a benchmarking methodology, and 2) 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 (Fig. 1, 2, Appendix 1), here we broadly classify models into one of five cohorts (Figure 2) based on the representation of urban impervious surfaces (buildings, roads etc):

(a) *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.

(b) *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).

(c) *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 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.

(d) *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 2D with an infinite canyon assumption. The details of canyon schemes vary widely, with single or multiple atmosphere layers, sub-facets (e.g. multi-faceted 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 (Fig. 1, 2).

(e) *More complex schemes [n=5]:* These resolve 3D interactions between urban facets using a variety of approaches. Repeated cuboids allow for two perpendicular streets while retaining some of the computational efficiency a canyon approach (Kanda et al., 2005). Statistical distributions can characterise realistic urban environments and have been used to determine 3D radiative interactions between buildings and urban vegetation similar to 3D radiative interactions between clouds (Hogan, 2019a, b). This allows complex urban environments to be simulated in a computationally efficient manner (Stretton et al., 2022). Building and tree resolving models represent 3D interactions more explicitly, allowing micro-climate 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 (Fig. 1). How models represent built, hydrological and anthropogenic processes are used here to obtain a measure of each model's 'total complexity' (Fig. 2). Participating model's parameterisations are individually summarised in Appendix 1.

Table 1: Participating models. Table 2 and Appendix 1 provide further details for each model. Section 2.1 provides an overview of urban modelling approaches and references.

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 Parameterization - 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)

Table 2: Participating model information. 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.

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

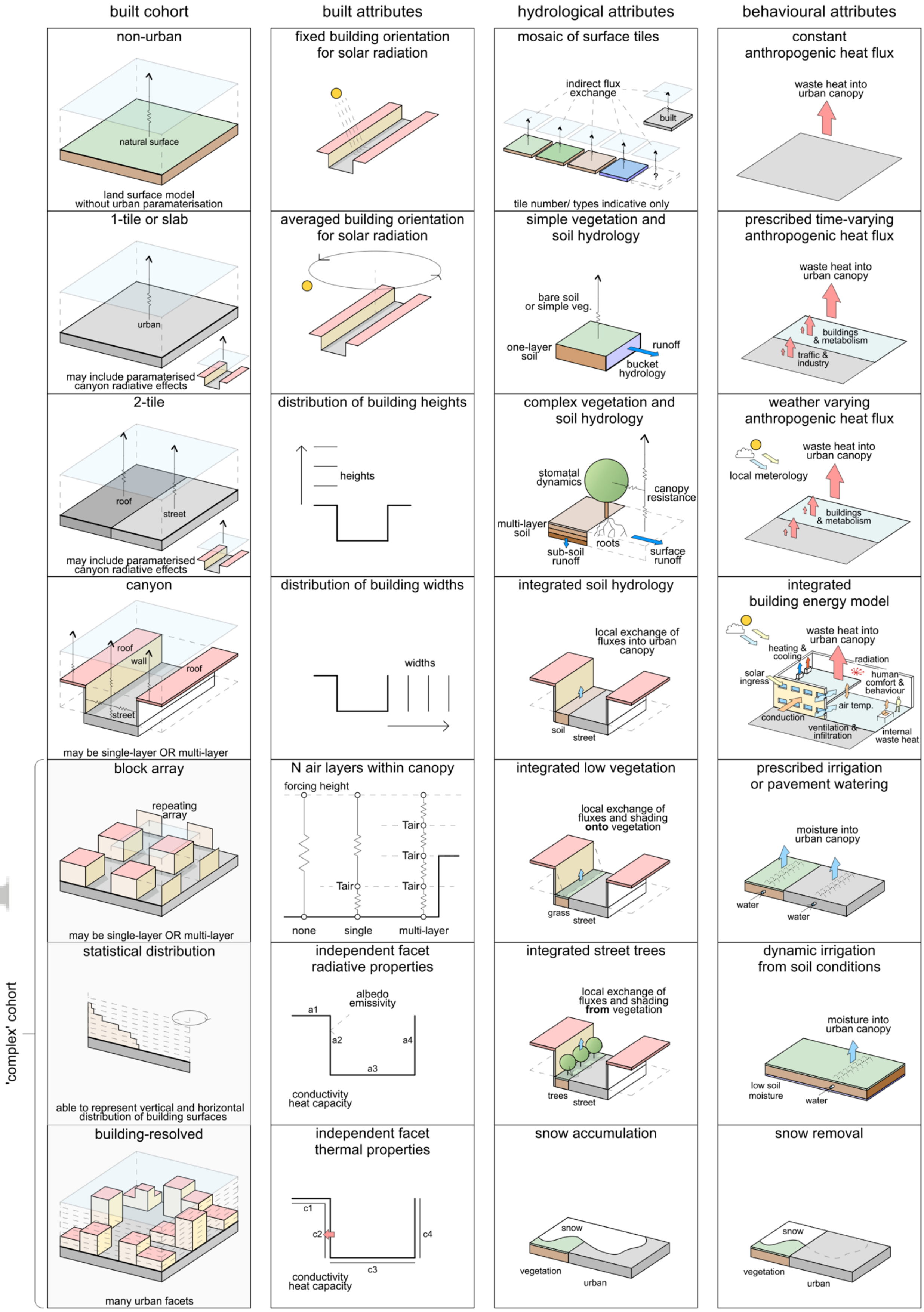
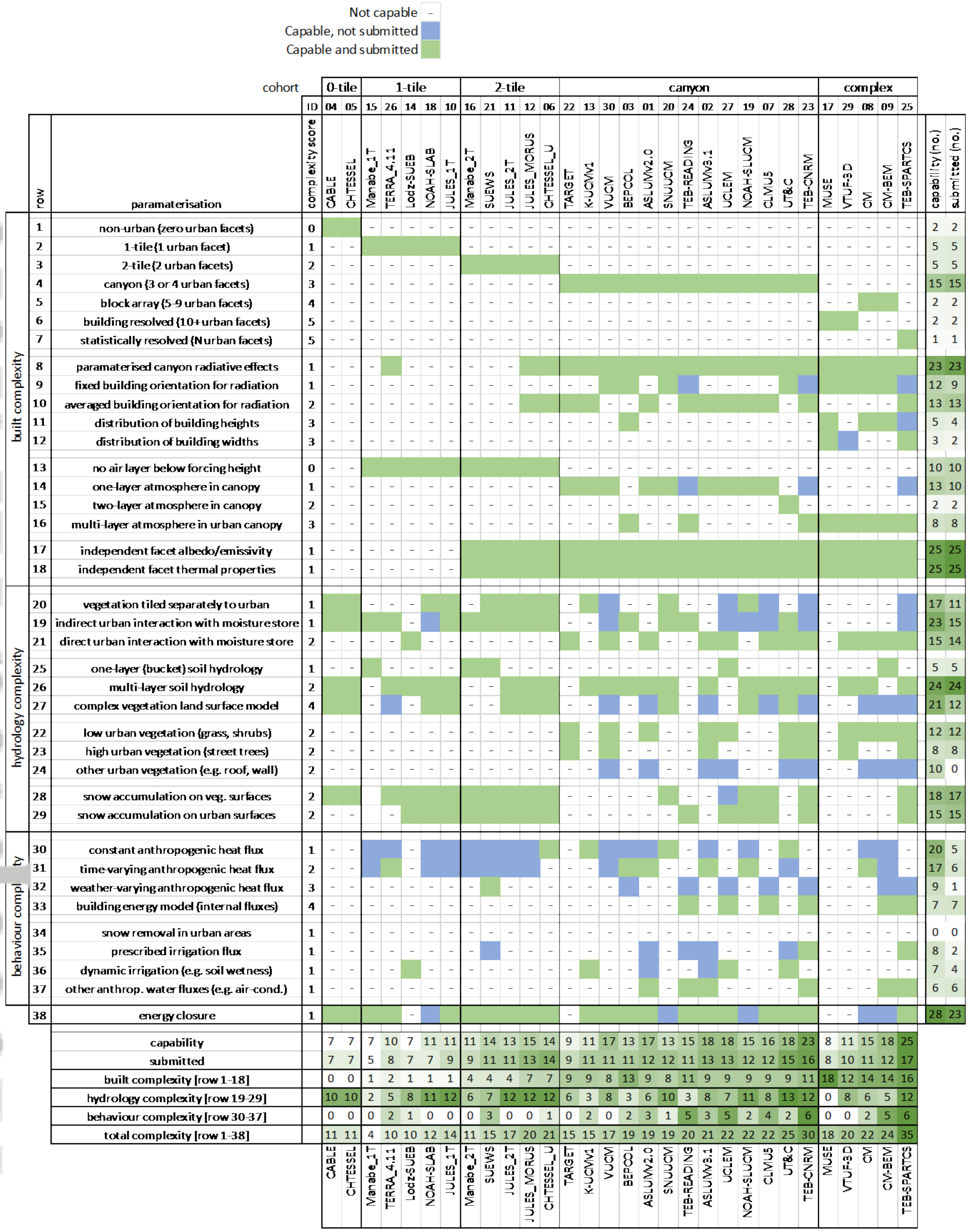


Figure 1: Model schematics of the main built, hydrological and behavioural attributes for participating models. Here models are categorised into five cohorts (left column) based on the geometric representation of buildings, with built, hydrological and behavioural attributes used to refine a 'total complexity'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



Not capable -
Capable, not submitted 🟦
Capable and submitted 🟩

			cohort	0-tile		1-tile					2-tile					canyon													complex						
			ID	04	05	15	26	14	18	10	16	21	11	12	06	22	13	30	03	01	20	24	02	27	19	07	28	23	17	29	08	09	25		
	row	parameterisation	complexity score	CABLE	CHTESSEL	Manabe_1T	TERRA_4.11	Lodz-SUEB	NOAH-SLAB	JULES_1T	Manabe_2T	SUEWS	JULES_2T	JULES_MORUS	CHTESSEL_U	TARGET	K-UCMv1	VUCM	BEPCOL	ASLUMv2.0	SNUUCM	TEB-READING	ASLUMv3.1	UCLEM	NOAH-SLUCM	CLMU5	UT&C	TEB-CNRM	MUSE	VTUF-3D	CM	CM-BEM	TEB-SPARTCS	capability (no.)	submitted (no.)
built complexity	1	non-urban (zero urban facets)	0	🟩	🟩	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	2	2
	2	1-tile (1 urban facet)	1	-	-	🟩	🟩	🟩	🟩	🟩	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5	5
	3	2-tile (2 urban facets)	2	-	-	-	-	-	-	-	🟩	🟩	🟩	🟩	🟩	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	5	5
	4	canyon (3 or 4 urban facets)	3	-	-	-	-	-	-	-	-	-	-	-	-	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	-	-	-	-	-	15	15
	5	block array (5-9 urban facets)	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	🟩	🟩	-	2	2
	6	building resolved (10+ urban facets)	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	🟩	🟩	-	-	-	2	2
	7	statistically resolved (N urban facets)	5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	🟩	1	1
	8	paramaterised canyon radiative effects	1	-	-	-	🟩	-	-	-	-	-	-	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	23	23
	9	fixed building orientation for radiation	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	🟩	🟩	-	🟩	🟦	-	-	-	-	🟩	🟦	🟩	🟩	🟩	🟩	🟦	12	9
	10	averaged building orientation for radiation	2	-	-	-	-	-	-	-	-	-	-	🟩	🟩	🟩	🟩	-	-	🟩	-	🟩	🟩	🟩	🟩	🟩	-	🟩	-	-	-	-	🟩	13	13
	11	distribution of building heights	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	🟩	-	-	-	-	-	-	-	-	-	🟩	-	🟩	🟩	🟦	5	4
	12	distribution of building widths	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	🟩	🟦	-	-	🟩	3	2
	13	no air layer below forcing height	0	-	-	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	10	10
	14	one-layer atmosphere in canopy	1	-	-	-	-	-	-	-	-	-	-	-	-	🟩	🟩	🟩	-	🟩	🟩	🟦	🟩	🟩	🟩	🟩	-	🟦	-	-	-	-	🟦	13	10
	15	two-layer atmosphere in canopy	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	🟩	-	-	-	-	-	-	2	2
	16	multi-layer atmosphere in urban canopy	3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	🟩	-	-	🟩	-	-	-	-	-	🟩	🟩	🟩	🟩	🟩	🟩	8	8
	17	independent facet albedo/emissivity	1	-	-	-	-	-	-	-	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	25	25
	18	independent facet thermal properties	1	-	-	-	-	-	-	-	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	25	25
hydrology complexity	20	vegetation tiled separately to urban	1	🟩	🟩	-	-	-	🟩	🟩	-	🟩	🟩	🟩	🟩	-	🟩	🟦	-	-	🟩	-	-	🟦	🟩	🟦	-	🟦	-	-	-	-	🟦	17	11
	19	indirect urban interaction with moisture store	1	🟩	🟩	🟩	🟩	-	🟦	🟩	🟩	🟩	🟩	🟩	🟩	-	-	🟦	🟩	-	🟩	🟩	-	🟦	🟦	🟦	🟩	🟦	-	-	-	-	🟦	23	15
	21	direct urban interaction with moisture store	2	-	-	-	-	🟩	-	-	-	-	-	-	-	🟩	-	🟩	-	🟩	-	-	🟩	🟩	-	🟩	🟩	🟩	-	🟩	🟩	🟩	🟩	15	14
	25	one-layer (bucket) soil hydrology	1	-	-	🟩	-	-	-	-	🟩	🟩	-	-	-	-	-	-	-	-	-	-	-	🟩	-	-	-	-	-	-	-	🟩	-	5	5
	26	multi-layer soil hydrology	2	🟩	🟩	-	🟩	🟩	🟩	🟩	-	-	🟩	🟩	🟩	-	🟩	🟩	🟩	🟩	🟩	-	🟩	-	🟩	🟩	🟩	🟩	-	🟩	🟩	-	🟩	24	24
	27	complex vegetation land surface model	4	🟩	🟩	-	🟦	-	🟩	🟩	-	-	🟩	🟩	🟩	-	-	🟦	-	🟦	🟩	-	🟦	-	🟩	🟦	🟩	🟦	-	-	🟦	🟦	🟦	21	12
	22	low urban vegetation (grass, shrubs)	2	-	-	-	-	-	-	-	-	-	-	-	-	🟩	-	🟩	-	🟩	-	-	🟩	🟩	-	-	🟩	🟩	-	🟩	🟩	🟩	🟩	12	12
	23	high urban vegetation (street trees)	2	-	-	-	-	-	-	-	-	-	-	-	-	🟩	-	🟩	-	-	-	-	🟩	-	-	-	🟩	🟩	-	🟩	-	-	🟩	8	8
	24	other urban vegetation (e.g. roof, wall)	2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	🟦	-	🟦	-	-	🟦	🟦	-	-	🟦	🟦	-	-	🟦	🟦	🟦	10	0
	28	snow accumulation on veg. surfaces	2	🟩	🟩		🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	-	-	-	-	-	🟩	-	-	🟦	🟩	🟩	-	🟩	-	-	-	-	🟩	18	17
	29	snow accumulation on urban surfaces	2	-	-		-	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	-	-	-	-	-	-	🟩	-	🟩	🟩	🟩	-	🟩	-	-	-	-	🟩	15	15
behaviour compl ty	30	constant anthropogenic heat flux	1	-	-	🟦	🟦	-	🟦	🟦	🟦	🟦	🟦	🟦	🟩	-	🟩	🟦	🟦	🟦	🟩	-	🟦	-	🟦	-	🟩	-	-	-	🟦	🟦	-	20	5
	31	time-varying anthropogenic heat flux	2	-	-	🟦	🟩	-	🟦	🟦	🟦	🟦	🟦	🟦	-	-	-	🟦	🟩	🟩	-	-	🟩	-	🟩	-	🟦	-	-	-	🟩	🟦	-	17	6
	32	weather-varying anthropogenic heat flux	3	-	-	-	-	-	-	-	-	🟩	-	-	-	-	-	-	🟦	-	-	🟦	-	🟦	-	🟦	-	🟦	-	-	-	🟦	🟦	9	1
	33	building energy model (internal fluxes)	4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	🟩	-	🟩	-	🟩	-	🟩	-	-	-	🟩	🟩	7	7
	34	snow removal in urban areas	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	0
	35	prescribed irrigation flux	1	-	-	-	-	-	-	-	-	🟦	-	-	-	-	-	-	-	🟦	-	🟦	🟦	-	-	-	🟦	🟩	-	-	-	-	🟩	8	2
	36	dynamic irrigation (e.g. soil wetness)	1	-	-	-	-	🟩	-	-	-	-	-	-	-	-	🟩	-	-	🟦	-	-	🟦	🟩	-	-	🟩	-	-	-	-	-	-	7	4
	37	other anthrop. water fluxes (e.g. air-cond.)	1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	🟩	-	🟩	🟩	-	-	-	-	🟩	-	-	-	🟩	🟩	6	6
	38	energy closure	1	🟩	🟩	🟩	🟩	-	🟦	🟩	🟩	🟩	🟩	🟩	🟩	-	🟩	🟩	🟩	🟩	🟦	🟩	🟩	🟩	🟦	🟩	🟩	🟩	-	-	🟦	🟦	🟩	28	23
		capability		7	7	7	10	7	11	11	11	14	13	15	14	9	11	17	13	17	13	15	18	18	15	16	18	23	8	11	15	18	25		
		submitted		7	7	5	8	7	7	9	9	11	11	13	14	9	11	11	11	12	12	11	13	13	12	12	15	16	8	10	11	12	17		
		built complexity [row 1-18]		0	0	1	2	1	1	1	4	4	4	7	7	9	9	8	13	9	8	11	9	9	9	9	9	11	18	12	14	14	16		
		hydrology complexity [row 19-29]		10	10	2	5	8	11	12	6	7	12	12	12	6	3	8	3	6	10	3	8	7	11	8	13	12	0	8	6	5	12		
		behaviour complexity [row 30-37]		0	0	0	2	1	0	0	0	3	0	0	1	0	2	0	2	3	1	5	3	5	2	4	2	6	0	0	2	5	6		
		total complexity [row 1-38]		11	11	4	10	10	12	14	11	15	17	20	21	15	15	17	19	19	19	20	21	22	22	22	25	30	18	20	22	24	35		
				CABLE	CHTESSEL	Manabe_1T	TERRA_4.11	Lodz-SUEB	NOAH-SLAB	JULES_1T	Manabe_2T	SUEWS	JULES_2T	JULES_MORUS	CHTESSEL_U	TARGET	K-UCMv1	VUCM	BEPCOL	ASLUMv2.0	SNUUCM	TEB-READING	ASLUMv3.1	UCLEM	NOAH-SLUCM	CLMU5	UT&C	TEB-CNRM	MUSE	VTUF-3D	CM	CM-BEM	TEB-SPARTCS		

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.

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 (Fig. 3). The neighbourhood is classed as an open low-rise (LCZ6) Local Climate Zone (Demuzere et al., 2022; Stewart and Oke, 2012). 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, b; Grimmond et al., 2011; Nice et al., 2018) or, when unavailable, estimated from high resolution global datasets (e.g. OpenLandMap soil datasets; Hengl, 2018a, b, c).

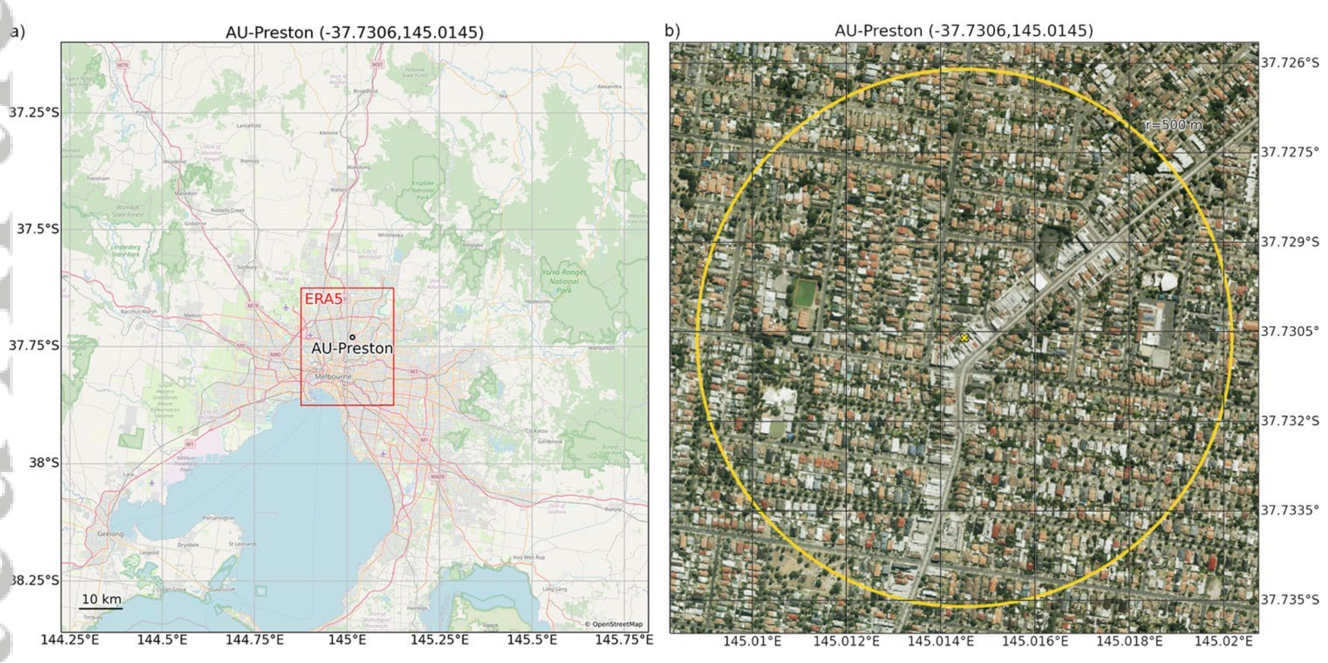


Figure 3: Study area, AU-Preston: (a) location within Melbourne, Australia with the extent of the ERA5 (Hersbach et al., 2020) grid cell used for gap-filling observations (red rectangle) (image: © OpenStreetMap contributors); and, (b) aerial imagery around the flux tower site (yellow cross with circle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

Table 3: Site-descriptive metadata. 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)*.

ID	Parameter	Value	Units	Footprint	Source
a)	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)
b)	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	2940	person km^{-2}	suburb average	(Coutts et al., 2007a)
21	Anthropogenic heat flux mean	11	W 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	1230	kg m^{-3}	250 m grid	(Hengl, 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 (2003-08-12 to 2004-11-28) at high frequency (1-10 Hz), which are then quality controlled and averaged to 30-minutes 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, b; Lipson et al., 2022a).

The site observations are split into (1) *forcing data:* provided to participants to drive models, and (2) *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 SWdown (except where SWdown=0 at night, which is assumed valid). SWup 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 hours) 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 (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).

Table 4: Observational data description and availability. 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).

Variable	Description	Units	Positive	All	Day	Night	DJF	MAM	JJA	SON
a. Forcing data				[%]	[%]	[%]	[%]	[%]	[%]	[%]
SWdown	Downward shortwave radiation	W m^{-2}	Downward	85.7	38.8	47.7	19.2	19.4	19.1	28.0
LWdown	Downward longwave radiation	W m^{-2}	Downward	71.8	38.2	33.6	19.2	19.4	13.7	19.5
Tair	Air temperature	K	-	100.0	52.3	47.7	19.2	19.4	23.5	37.9

Qair	Specific humidity	kg kg^{-1}	-	100.0	52.3	47.7	19.2	19.4	23.5	37.9
PSurf	Station air pressure	Pa	-	86.3	46.0	40.2	19.1	19.1	16.4	31.7
Wind_N	Northward wind component	m s^{-1}	Northward	99.9	52.2	47.7	19.2	19.4	23.5	37.9
Wind_E	Eastward wind component	m s^{-1}	Eastward	98.9	51.7	47.3	18.8	19.4	23.4	37.2
Rainf	Rainfall rate	kg m^{-2} s^{-1}	Downward	100.0	52.3	47.7	19.2	19.4	23.5	37.9
Snowf	Snowfall rate	kg m^{-2} s^{-1}	Downward	0.0	0.0	0.0	0.0	0.0	0.0	0.0
b. Analysis data										
SWup	Upward shortwave radiation	W m^{-2}	Upward	35.9	35.9	0.0	11.5	7.3	5.9	11.2
LWup	Upward longwave radiation	W m^{-2}	Upward	66.2	35.1	31.1	19.0	15.8	13.4	18.0
Qle	Latent heat flux	W m^{-2}	Upward	43.1	19.8	23.4	11.3	9.3	11.5	10.9
Qh	Sensible heat flux	W m^{-2}	Upward	43.3	19.8	23.5	11.3	9.4	11.6	11.0
Qtau	Momentum flux	N m^{-2}	Downward	73.3	35.4	37.9	18.9	15.8	14.1	24.5

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 inter-annual 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 timeseries for ten 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, 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 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 shortwave (SWup) and longwave (LWup) radiation, sensible (Qh) and latent heat flux (Qle)—as well as the momentum flux (Qtau)

(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 timesteps).

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 (Qanth) and water (Qirrig) fluxes, storage heat flux (QStor), roof, wall and road surface temperatures (RoofSurfT, WallSurfT, RoadSurfT), bulk air temperature within buildings and in street canyons (TairBuilding, TairCanyon), and urban canopy albedo (UAlbedo). 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, timeseries and energy closure plots are provided to participants by project coordinators. Shortwave 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 shortwave radiation dominates the energy balance (Best and Grimmond, 2015).

Following feedback, participants had an opportunity to resubmit prior to more complete analysis and final error statistics being shared. The number of submissions (including the first) varied from 1 to 9. Initial checks identified human errors including incorrect start times, mislabelling of outputs, variable sign errors, forcing interpolation errors (where model timestep 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 timeseries plots of all submitted variables and individual model results and metadata (archived in supplementary material).

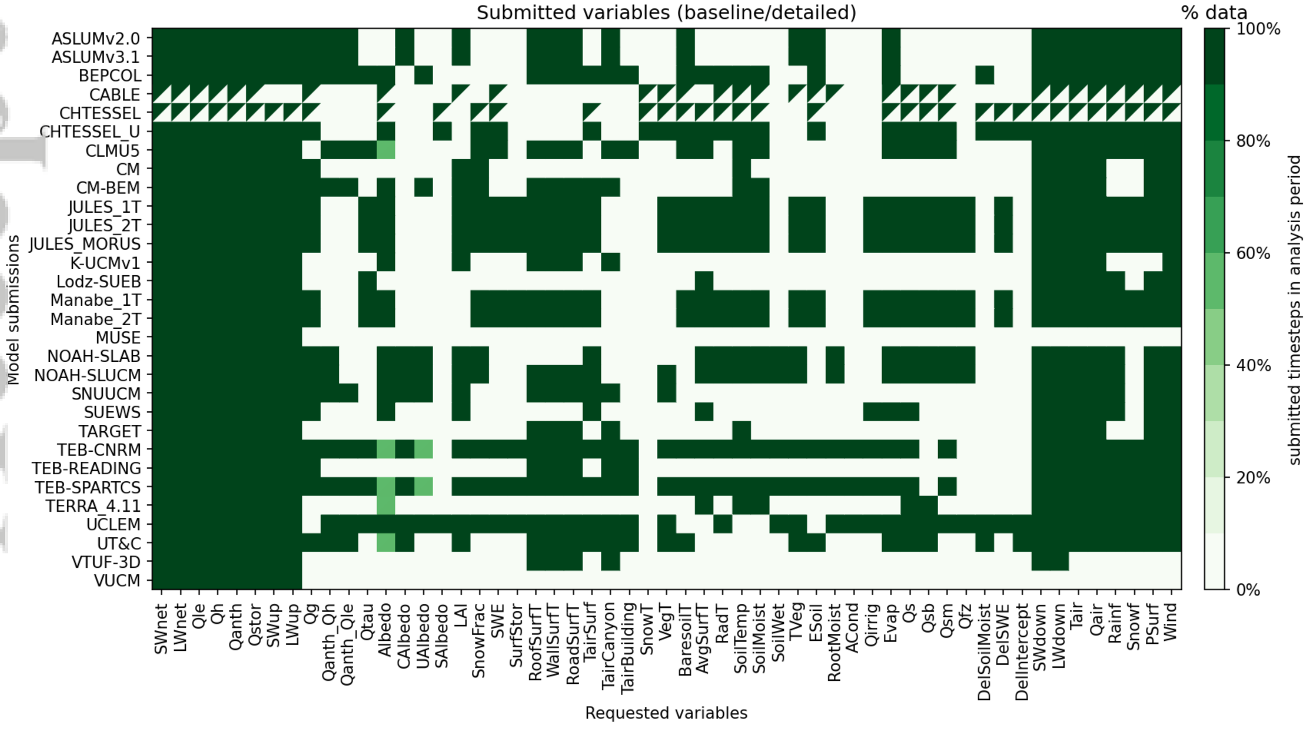


Figure 4: Submitted variables. Variables analysed here are defined in Table 4, with others defined in the Urban-PLUMBER modelling protocol (Lipson et al., 2020), following the ALMA naming conventions (Bowling and Polcher, 2001). Two models (CABLE and CHTESSEL) submitted only the baseline experiment, 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 timestep 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:		
SWdown	ensure timestamps of submission matched expectations	some
SWdown (subset)	check if modelled SWdown matches forcing: some using < 30-min timesteps 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

2.3.Evaluation methods

2.3.1. Benchmarks

Following PLUMBER (Best et al., 2015), we use benchmarks (Fig. 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 shortwave radiation (SWdown), air temperature (Tair), and relative humidity (RH) from 20 urban sites (Table A.2). 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:

a) **Manabe_1T:** A simple 'slab and bucket' model (Fig. 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 evaporate (Manabe, 1969). At each timestep, 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:

b) **REG1-SWdown:** Linear regression with one variable (SWdown, Fig. 5b) is used separately to predict SWup, LWup, Qh, Qle, and Qtau. At night all predicted values are constant because SWdown = 0 W m^{-2}.

c) **REG2-SWdown-Tair:** Two-variable (SWdown and Tair) linear regression (Fig. 5c) provides some information at night and provides benefit for variables strongly dependent on temperature (e.g. LWup and Qh).

d) **KM3-SWdown-Tair-RH:** A piecewise multi-variable regression. Following PLUMBER's conceptual arguments, three predictor variables (SWdown, Tair 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 (Fig. 5d). To use this benchmark, at each timestep 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:

e) **KM3-IS-SWdown-Tair-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.

f) **KM4-IS-SWdown-Tair-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).

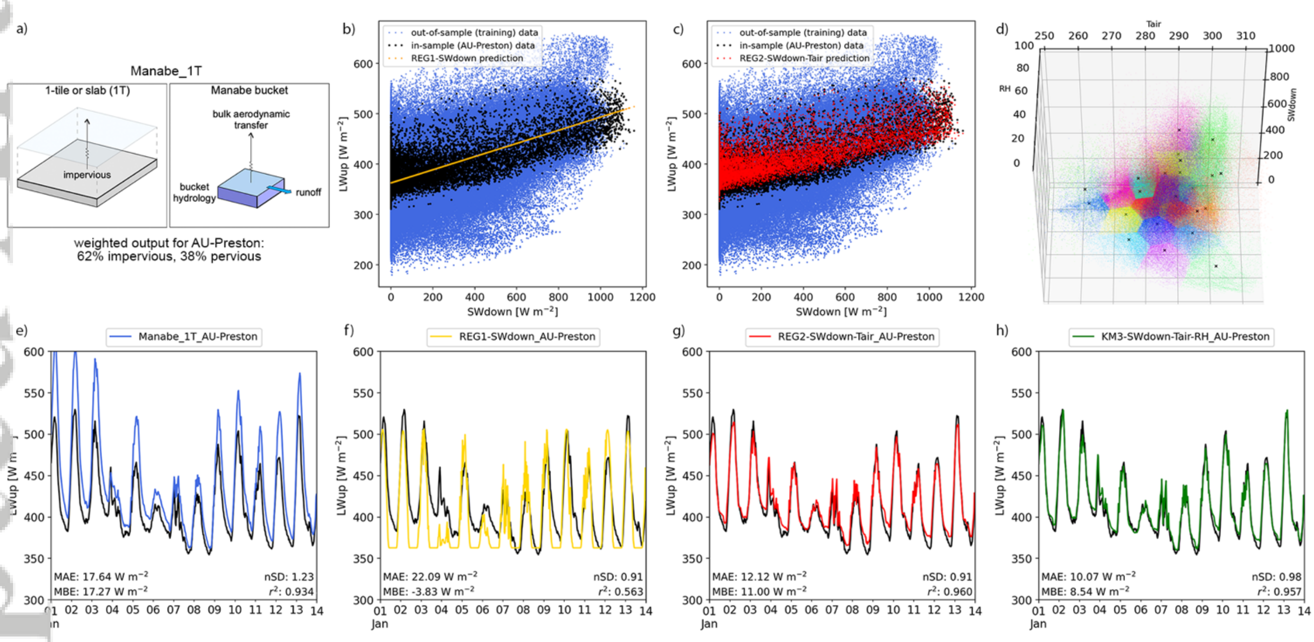


Figure 5: Example benchmark diagrams. **a**) Manabe_1T, **b**) REG1-SWdown, **c**) REG2-SWdown-Tair, **d**) KM3-SWdown-Tair-RH with cluster centroids (black crosses), and (**e-h**) predictions using corresponding (a-d) benchmarks. A two-week period of upward longwave radiation (LWup) is used as an example 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 (r²), all defined in Table A.1.

2.3.2. Error metrics

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

1) *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.
2) *Extremes* of the observed distribution: absolute error at the 5th and 95th percentile of observed and modelled outputs.
3) *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 A.1). 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 A.1.

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 ith model using the minimum and maximum metric result across all models and benchmarks for that variable:

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

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:

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

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: Fig. 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 m^{-2}). Subsequently, three statistics (for correlation, variance and difference errors) are analysed in a Taylor diagram (Taylor, 2001) (Section 3.2: Fig. 7). Aggregated benchmark performance scores (Section 2.3.3) are then analysed in benchmarking diagrams, using common error metrics (Section 3.3: Fig. 8). Finally, all metrics (common, extreme and distribution) are used to compare models with benchmarks (Section 3.4: Fig. 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 (MAE)

Individual model MAE results are combined into boxplots (Fig. 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 timestep; rightmost column) and the benchmarks (coloured horizontal lines) are also shown (Fig. 6).

For upward shortwave radiation (SWup), 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 m^{--2}) indicates this flux can be well simulated with few inputs.

For upward longwave radiation (LWup), 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 timeseries outperforms the physical and out-of-sample benchmarks. The LWup benchmark MAE values are larger (5.2 — 22.4 W m^{-2}) than for SWup, indicating the flux is more challenging to predict with the information available to benchmarks.

For sensible heat flux (Qh), 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 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 m^{-2}) indicates that the flux is more challenging to predict than either radiative flux.

For latent heat flux (Qle), 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. Qle has a relatively high benchmark range (18.6 – 26.1 W 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 SWup, Qh and Qle, but not for LWup (t-test, p<0.05). The range of MAE for the detailed simulations are generally smaller than for the baseline, indicating models benefited from additional site information. The difference in the mean MAE for baseline and detailed experiments reaches significance in SWup only.

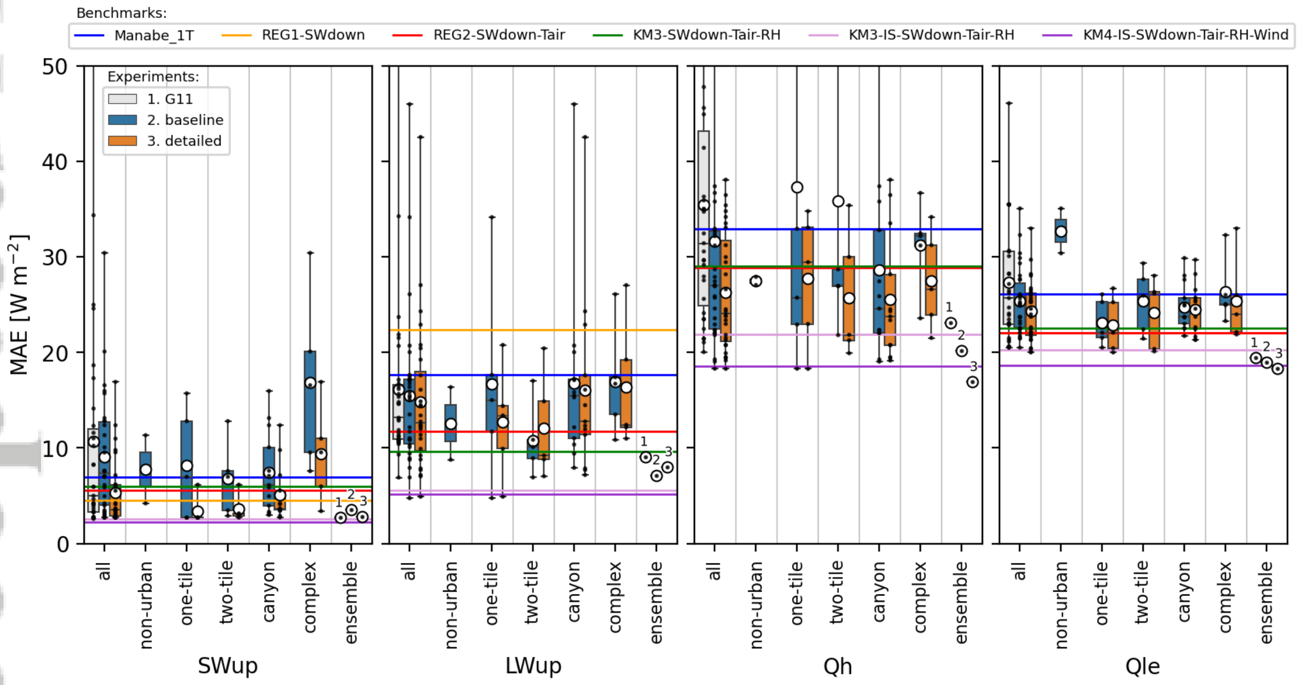


Figure 6: Mean absolute error (MAE) boxplot results. Individual models (dots) are split into cohorts (Section 2.1) with benchmarks (horizontal lines) for four fluxes: upward shortwave radiation (SWup), upward longwave radiation (LWup), sensible heat flux (Qh) and latent heat flux (Qle). Boxplots are 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: 1) a difference error metric (centred root mean square error: cRMSE); 2) a variance error metric (modelled standard deviation normalised by observed standard deviation: $\widetilde{\sigma}$) and 3) a correlation error metric (Pearson's correlation coefficient: r; all defined in Table A. 1). 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 (Fig. 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 SWup, most Urban-PLUMBER (UP) models and benchmarks are grouped tightly around the observation star (Fig. 7a). Some UP models (e.g. 20, 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 SWup (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 (Fig. 7).

For LWup (Fig 7b), many participating models have larger variance than observed because of an overprediction in the diurnal range of LWup. 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 LWup 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 Qh (Fig 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. 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 Qh values or a general overestimation of the variability in Qh. As cRMSE is "centred" it does not measure bias error (Taylor, 2001). For Urban-PLUMBER, the MBE for Qh 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 Qh.

Compared with other fluxes, the poorer cRMSE of the six benchmarks for Qle (Fig7d) 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 Qle output (i.e. some assumed Qle = 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 Qle). Combined with the improved ensemble mean Qh MBE, this indicates the UP models are better at partitioning available energy into Qh and Qle. No model in either project outperformed in-sample empirical benchmarks for Qle (Table 6).

Eleven models provided the simulated momentum flux (Qtau) from both Urban-PLUMBER and G11 (Fig 7e). Performance in both projects are 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 3-variable and 4-variable in-sample benchmarks, and all were able to beat out-of-sample empirical benchmarks.

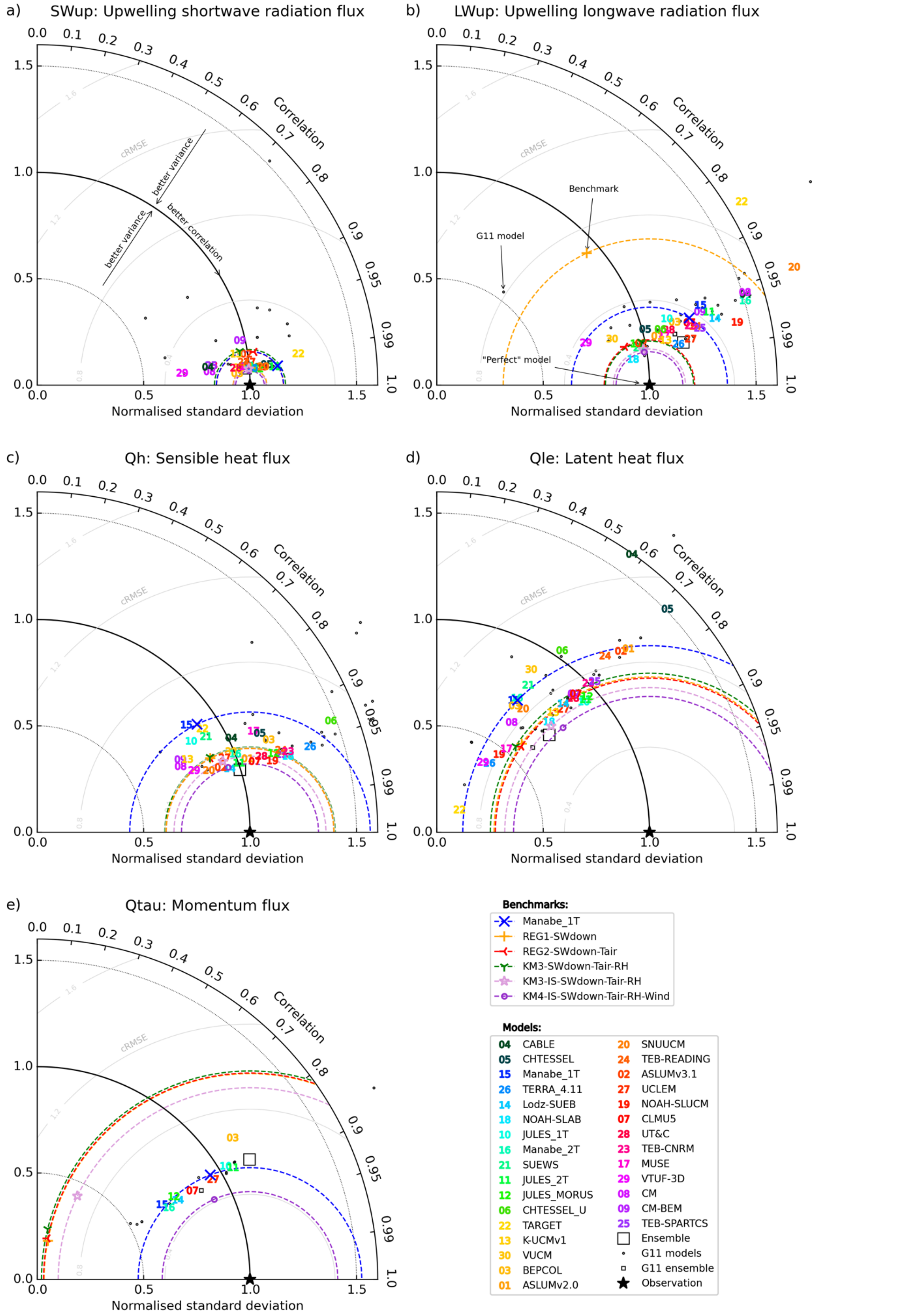


Figure 7: Taylor diagram combines the normalised standard deviation ($\tilde{\sigma}$), correlation coefficient (r) and cRMSE (defined Table A.1). Models (coloured numbers) have better performance if closer to star at diagram base, with cohort colours: non-urban (dark greens), one-tile (blues), two-tile (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).

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

	Flux	SWup		LWup		Qh		Qle		Qtau	
	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

3.3. Benchmarking evaluation: common metrics

We can evaluate model results relative to the benchmarks for various experiments using the aggregated scores (Eq. 1 and 2) from four common metrics (MAE, MBE, nSD, R). A lower relative score indicates better performance (Fig. 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' (Fig. 2), which includes hydrology and anthropogenic related characteristics.

For SWup, 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 ten 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 SWup (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 LWup, 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 LWup performance by utilising the detailed building morphology information provided (e.g. representing canyon longwave 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 Qh, 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). Qh is the only variable where the ensemble mean beats all the benchmarks for common metrics.

Although additional site information degrades Qle 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 Qle magnitudes. The three best performing models in Qle (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 Qle, 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 Qtau 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.

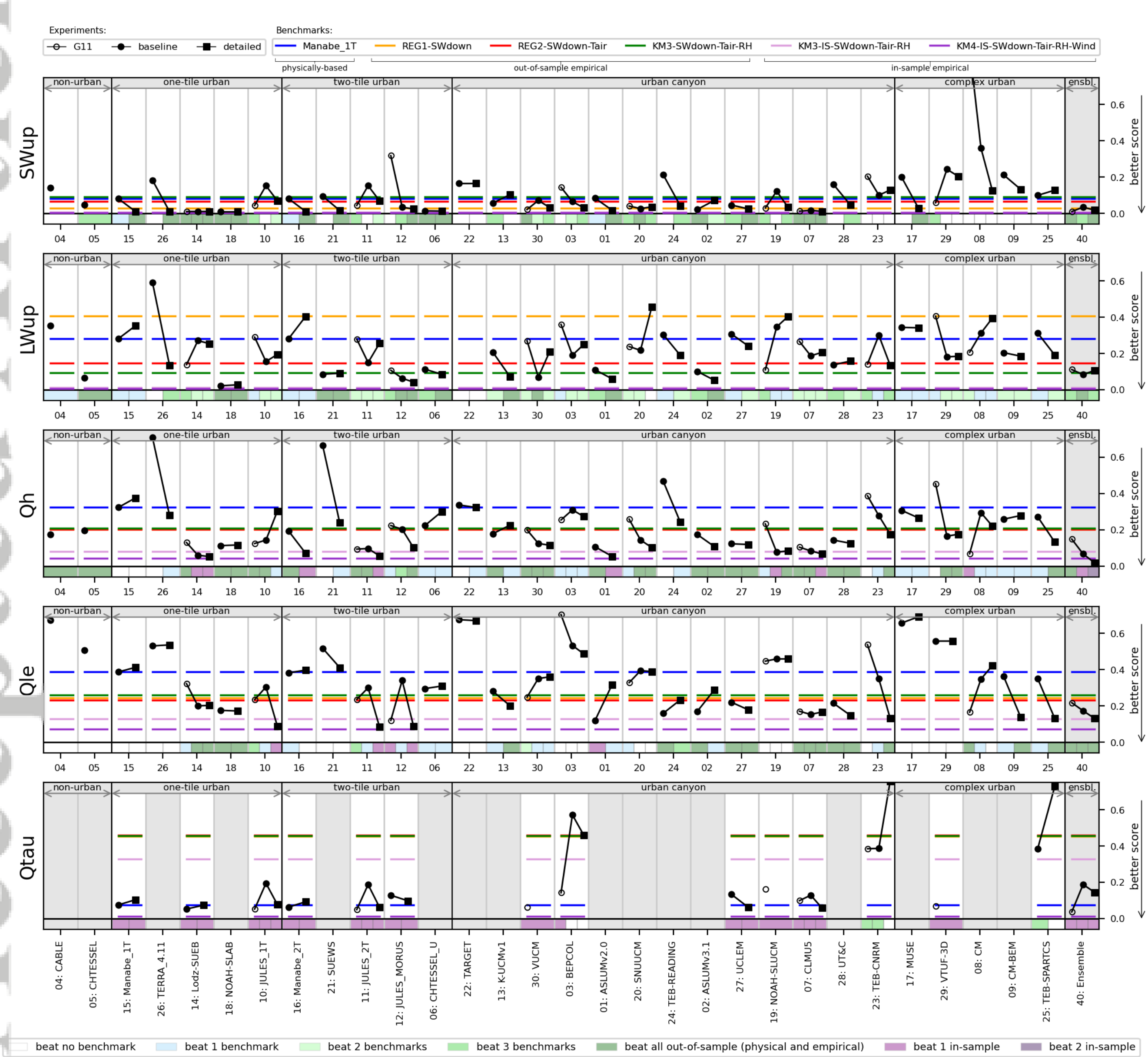


Figure 8: Benchmarking assessment for 'common' set of metrics (MAE, MBE, nSD, R) showing benchmarks (coloured lines) and models (black markers). Lower scores are better (Eq. 1 – 2). Model columns include up to three markers for different experiment submissions (1. PILPS-Urban Phase 2 Stage 4 (G11), 2. 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 (Fig. 2). For Qtau, grey indicates no submission.

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 Fig. 8 with “extreme” and “distribution” error metrics in Fig. 9. Relative performances (as in Fig. 8) for metric groups are provided in supplementary material (Fig. S1 – S3).

For Qh and Qle, 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 Qh challenging (white or blue) because they overestimate the upper Qh tail. For Qtau, 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-SWdown-Tair-Rh-Wind). Most models assessed for Qtau perform similarly to the simple physically-based benchmark (Fig. 8).

The ensemble mean timeseries performs strikingly well across all fluxes, beating in-sample benchmarks for many metrics in Qh, Qle and Qtau, and beating most out-of-sample benchmarks in SWup and LWup. 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.

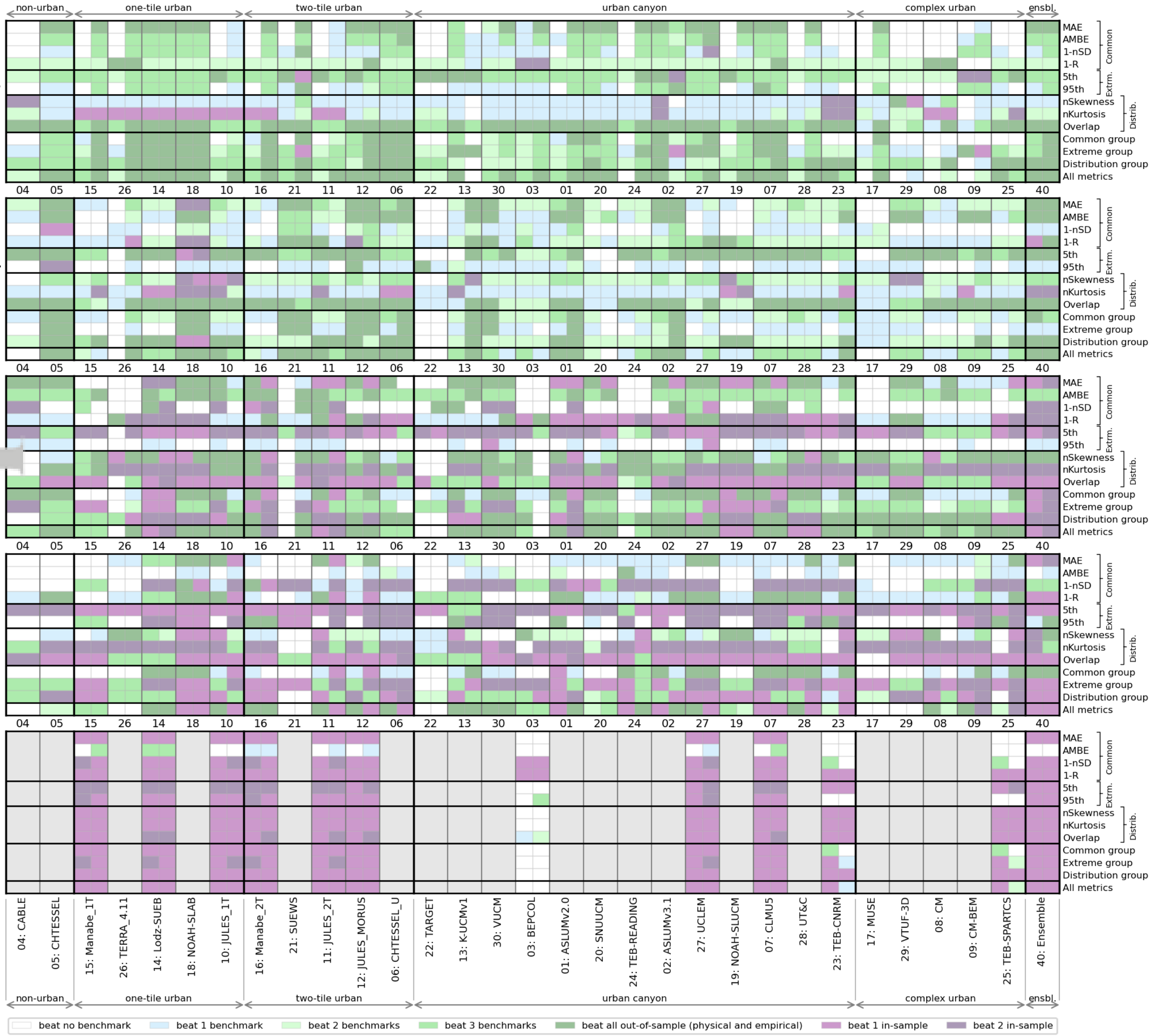


Figure 9: All metric benchmarking results. Models are ordered nominally by increasing complexity (Fig. 2) showing benchmark scorecard of individual and aggregated metrics based on ability to outperform benchmarks (colour) for both the baseline (left) and detailed (right) simulations. Note a single box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

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 (Fig. 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 LWup). For cRMSE (Table 6), we find broad improvement for upward shortwave (SWup), sensible (Qh) and latent (Qle) heat fluxes, but little or no improvement in upward longwave (LWup) and momentum (Qtau) fluxes. When assessing performance using four common evaluation metrics (Fig. 8), the ensemble mean has clearly improved for Qh and Qle but is little changed or slightly degraded for SWup, LWup and Qtau.

These results suggest the current generation of models is performing better than the G11 models for Qh and Qle. 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 (Fig. 7) indicates time-of-day errors (i.e. human rather than model errors). However, when all models with SWup correlations below 0.99 are excluded from the analysis (i.e. retaining only those with good timing) , G11 models still perform poorer for LWup, Qh and Qle compared with the current cohort (Fig. S4), implying other factors have led to performance improvements.

Feedback provided to participants prior to final evaluation (Table 5) significantly reduces SWup and LWup errors for some models, and to a lesser degree also reduced Qh and Qle errors (Fig. 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 (Fig 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 which undertook the extended 10-year spinup tend to have better performance in detailed simulations than those that did not (Fig. S7), however this effect was small. Models with a closed surface energy budget have better performance in radiative fluxes, but not in turbulent fluxes (Fig. S8). Fully separating the various influences (better models, more experienced modellers, better spinup 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 (Fig. 8), daytime only (Fig. S9) and nighttime only (Fig. S10) results are presented in supplementary material. For LWup 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 in-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 longwave trapping between urban structures that simpler schemes could not. In this project, some models in every cohort perform well for LWup 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 Qh, Qle, and Qtau 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 Qh, 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 Qle appears to be poorly modelled according to the Taylor plot statistics (Fig. 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 Qle errors than, for example, SWup errors (Fig. 7). Benchmark assessment of extremes and distribution skill finds Qle to be one of the better modelled fluxes, with models able to outperform the in-sample empirical benchmarks in many instances (Fig. 9: purple). Likewise, Qh is well modelled for the common (Fig. 8) and other (Fig. 9) metrics, compared to benchmarks. Fewer models outperform benchmarks for LWup (Fig. 7, 8), particularly in the daytime (Fig. S9). More site information generally degraded the ensemble mean skill in LWup (Fig. 6, 8), except for more complex models at night (Fig. S10). The overall poor performance in LWup compared with benchmarks indicates an area for which model development may prove beneficial. This may be difficult, as LWup 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 LWup is additionally complicated by the fact that the footprint of radiative observations from a flux tower differ from the footprint of the turbulent fluxes (Sailor, 2011; Schmid et al., 1991), 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 (Fig. 6, 8, 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' representation have the simplest soil, water and vegetation approaches (Fig. 2). The simpler models within a cohort (i.e. left side of each cohort, Fig. 8, 9) often had poorer intra-cohort 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 Qle, performing worse than benchmarks and all other model cohorts in this flux (Fig. 6, 7, 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 Qh and Qle, and in some cases for the three or four variable in-sample benchmarks (Fig. 6 - 9). For common metrics, PLUMBER (Best et al., 2015) found no model able to outperform a single-variable linear regression using SWdown for Qh, or a three-variable regression for Qle 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-SWdown-Tair-RH) performed poorer than simpler regressions for Qh and Qle (Fig. 6, 7, 8), indicating overfitting. However, some models outperform the regressions trained in-sample (i.e. using only AU-Preston data), 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 multi-site evaluation is required to confirm these initial results (now underway).

In Urban-PLUMBER, we focus on bulk local scale surface-atmosphere exchanges as these variables and scale acts 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 model intercomparison projects 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 focused on SWup to check that forcing and output timing aligned, and we link this to the net improvement in SWup performance seen (cf. G11, Fig. 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 1) aspects of model design, 2) user model configuration, or 3) model assessment methods (e.g. variables, metrics, spatial and time scales). A multi-site 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 shortwave radiation (SWup), sensible (Qh) and latent (Qle) heat fluxes, but little/ no improvement in upward longwave radiation (LWup) and momentum (Qtau) fluxes.
- As in PILPS-Urban, the ensemble mean timeseries 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 Qle, 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 LWup, thus indicating an area for which future model development may prove beneficial.
- Many models outperform the non-linear three-variable empirical benchmarks for Qh, 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 SWdown 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 shortwave 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.

6. Acknowledgements

The project's coordinating team are 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 and National Computational Infrastructure (NCI) Australia. G.J. Steeneveld and A. Tsiringakis acknowledges 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 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).

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.

7. Author contributions

M. Lipson, S. Grimmond, M. Best, G. Abramowitz: project conceptualisation and methodology. M. Lipson: project coordination, data curation, software development, validation, analysis, investigation, visualisation. M. Lipson with S. Grimmond: original draft manuscript. A. Pitman: Supervision and resources. A. Coutts and N. Tapper: collection of observations used in this evaluation. All other co-authors (listed alphabetically): investigation (development, creation and submission of model data), review and editing of the manuscript.

8. Supplementary material

Supplementary material include:

- Fig. S1: Benchmark evaluation for extreme group of metrics (equivalent to Fig. 8).
- Fig. S2: Benchmark evaluation for distribution group of metrics.
- Fig. S3: Benchmark evaluation for all metrics combined.
- Fig. S4: Categorical MAE plot considering only models with SWup correlations of > 0.99 to exclude models with time-of-day errors.

- Fig. S5: Comparison of MAE for initial and baseline submissions.
- Fig. S6: Categorical MAE plot for participants with/without previous experience in modelling the AU-Preston site.
- Fig. S7: Categorical MAE plot for models that did/did not complete full the spin-up period.
- Fig. S8: Categorical MAE plot for models that did/did not close the surface energy balance.
- Fig. S9: Benchmark evaluation for daytime periods (equivalent to Fig. 8)
- Fig. S10: Benchmark evaluation for nighttime periods (equivalent to Fig. 8)

Associated results include:

- Individual model results (error metrics, individual timeseries and time-averaged plots, variable distribution plots, submitted metadata, and list of values outside of ALMA expected ranges.
- Collective timeseries for every submitted output in the baseline and detailed experiments

9. Data availability

Supplementary material 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 timeseries data are openly available from https://doi.org/10.5281/zenodo.7104984 (Lipson et al., 2022c).

Benchmark timeseries data are available from https://doi.org/10.5281/zenodo.7330052 (Lipson and Best, 2022).

Appendix A1

A1.1 Error metric definitions

Table A.1: Error statistics and metrics used to evaluate models. 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. $\bar{O}$) indicates the mean of n samples. n varies with observational availability. X_5 is value of X at the 5th 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).

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	$\frac{\sum(M_i - \bar{M})(O_i - \bar{O})}{\sqrt{\sum(M_i - \bar{M})^2}\sqrt{\sum(O_i - \bar{O})^2}}$	-
Standard deviation	σ$_X$	$\sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$	-
Normalised standard deviation	$\tilde{\sigma}$	$\frac{\sigma_M}{\sigma_O}$	(Taylor, 2001)
Skewness	μ_X	$\frac{1}{n}\sum\left(\frac{X_i - \bar{X}}{\sigma_X}\right)^3$	(Best et al., 2015)
Kurtosis	K_X	$\frac{1}{n}\sum\left(\frac{X_i - \bar{X}}{\sigma_X}\right)^4 - 3$	(Best et al., 2015)
Perkins skill score	PSS	$\sum_1^{100} \min(bin_{M,k}, bin_{O,k})$	(Perkins et al., 2007)
Centred and normalised root 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}	$\lvert MBE\rvert$	(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 5th 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	$m_{overlap}$	$1-\mathrm{PSS}$	(Haughton et al., 2017)

A1.2 Benchmark observational data

Table A.2: Site locations of tower observational data used to generate the empirical benchmarks for this study. 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., 2022c, a).

Sitename	City	Country	Observed period	Latitude	Longitude	References
AU-Preston	Melbourne	Australia	Aug 2003 – Nov 2004	-37.7306	145.0145	(Coutts et al., 2007a, b)
AU-SurreyHills	Melbourne	Australia	Feb 2004 – Jul 2004	-37.8265	145.099	(Coutts et al., 2007a, b)
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	(Järvi et al., 2018; Nordbo et al., 2013)
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	(Jo et al., n.d.; Hong et al., 2020)
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	(Fortuniak et al., 2013; Pawlak et al., 2011)
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	(Bjorkegren et al., 2015; Kotthaus and Grimmond, 2014a, b)
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, 2017; Chow et al., 2014)

A1.3 Participating model descriptions

For each model (Figure A.1- A.30) 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.

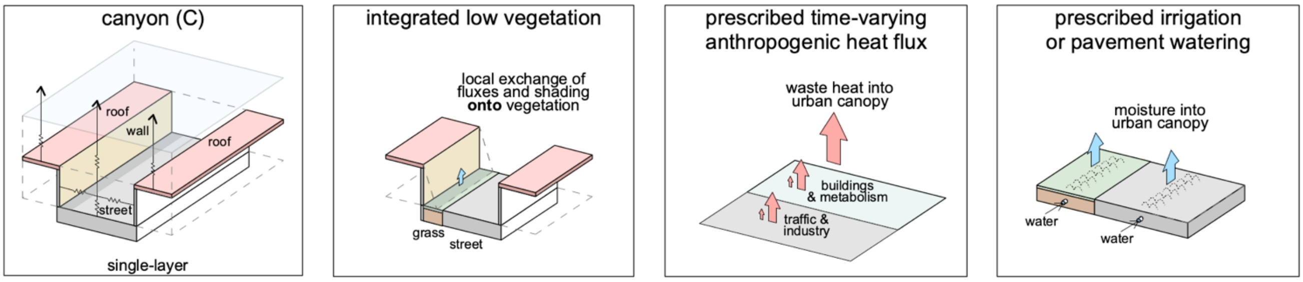

canyon (C)
roof
wall
roof
street
single-layer
integrated low vegetation
local exchange of fluxes and shading onto vegetation
grass
street
prescribed time-varying anthropogenic heat flux
waste heat into urban canopy
buildings & metabolism
traffic & industry
prescribed irrigation or pavement watering
moisture into urban canopy
water
water

Figure A.1: 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 sub-facet heterogeneity and urban vegetation. ASLUMv2 incorporates detailed hydrology, multi-layer soil/ground and roof vegetation (via a multi-layer green roof model), and enables simulations of irrigation, anthropogenic heat, and urban oasis effect (Yang et al., 2015).

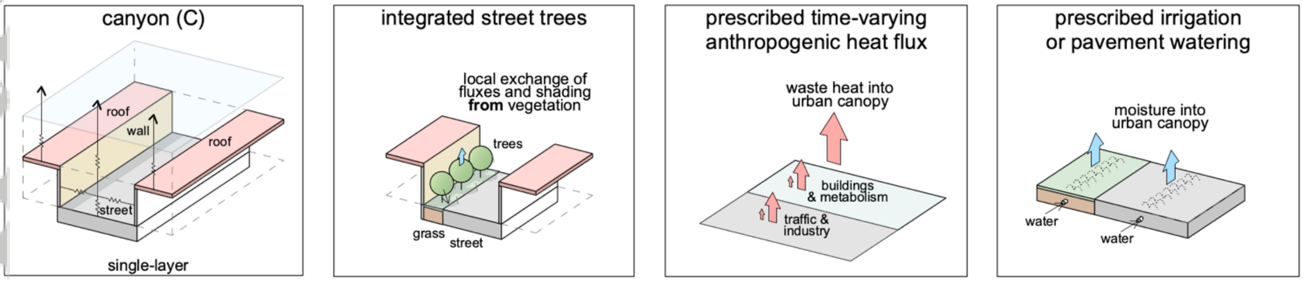

canyon (C)
roof
wall
roof
street
single-layer
integrated street trees
local exchange of fluxes and shading from vegetation
trees
grass
street
prescribed time-varying anthropogenic heat flux
waste heat into urban canopy
buildings & metabolism
traffic & industry
prescribed irrigation or pavement watering
moisture into urban canopy
water
water

Figure A.2: ASLUMv3.1 (Wang et al., 2021, 2013) is the same as ASLUMv2.0 (Fig. A.1) 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).

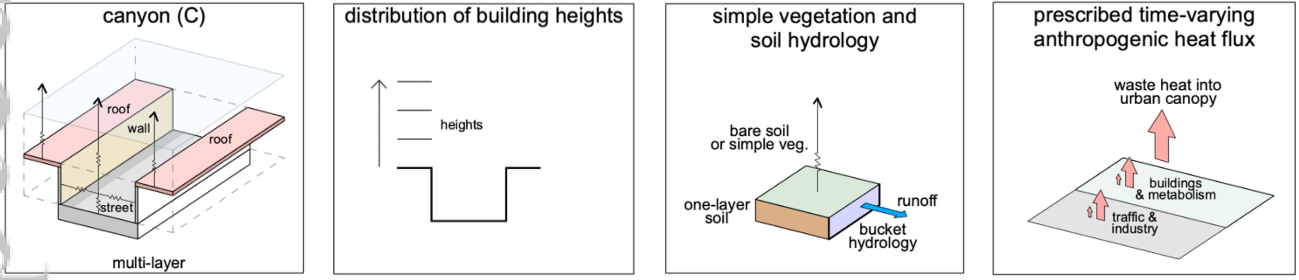


Figure A.3: BEPCOL is a multilayer canyon model based on the building effect parameterization (BEP; Martilli et al., 2002) with parameterizations for drag coefficient and the length scales used for turbulent transport and turbulence dissipation (Simón-Moral et al., 2017). BEPCOL does not explicitly consider vegetation and the non-urban fraction is computed by the bare soil model from RAMS (Tremback and Kessler, 1985).

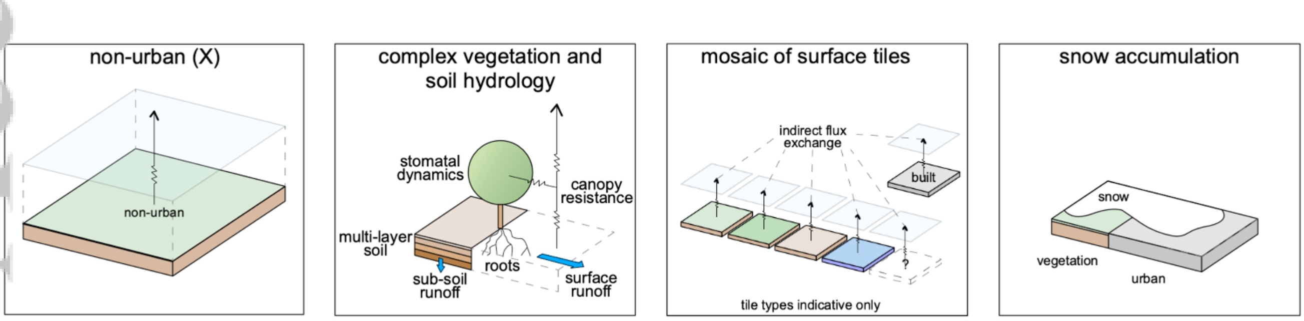


Figure A.4: CABLE (Community Atmosphere–Biosphere Land Exchange model) (Kowalczyk et al., 2006; Wang et al., 2011a) is used in regional and global climate models including ACCESS (Kowalczyk et al., 2013). CABLE has a one-layer, two-leaf canopy vegetation scheme with up to five tiles (vegetation types, 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.

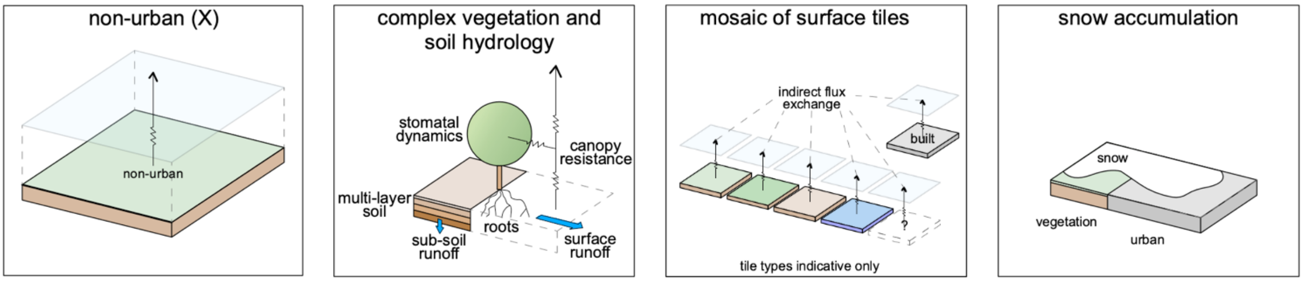


Figure A.5 CHTESSEL (Carbon-Hydrology Tiled ECMWF Forecasts Scheme for Surface Exchanges over Land) is the land surface model used in the Integrated Forecast System (IFS). This is used by ECMWF for weather forecast and to create reanalysis products (Balsamo et al., 2009; Boussetta et al., 2013; ECMWF, 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.

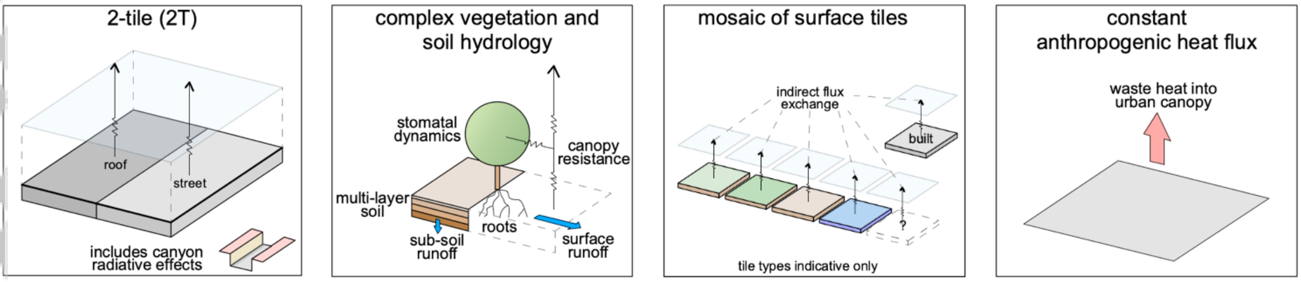

2-tile (2T)
roof
street
includes canyon radiative effects
complex vegetation and soil hydrology
stomatal dynamics
canopy resistance
multi-layer soil
roots
sub-soil runoff
surface runoff
mosaic of surface tiles
indirect flux exchange
built
?
tile types indicative only
constant anthropogenic heat flux
waste heat into urban canopy

Figure A.6: 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 (Fig. A.5). It follows MORUSES (Fig. A.13) infinite canyon assumptions for radiative effects. The urban surfaces (cf. CHTESSEL) have increased runoff and reduced soil infiltration.

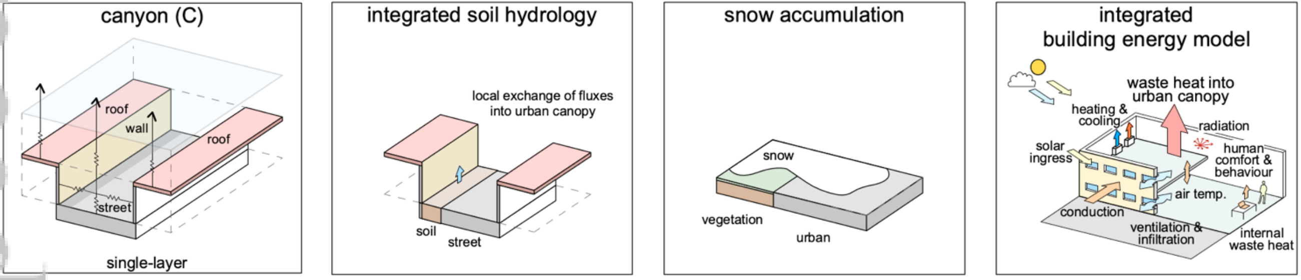

canyon (C)
roof
wall
roof
street
single-layer
integrated soil hydrology
local exchange of fluxes into urban canopy
soil
street
snow accumulation
snow
vegetation
urban
integrated building energy model
waste heat into urban canopy
heating & cooling
radiation
solar ingress
human comfort & behaviour
air temp.
conduction
ventilation & infiltration
internal waste heat

Figure A.7: 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 and Feddema, 2020; Oleson et al., 2010). 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 shortwave and longwave 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).

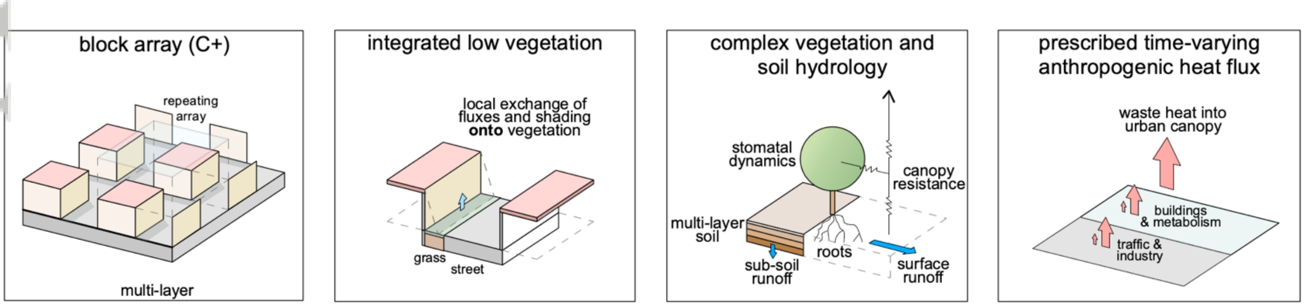

block array (C+)
repeating array
multi-layer
integrated low vegetation
local exchange of fluxes and shading onto vegetation
grass
street
complex vegetation and soil hydrology
stomatal dynamics
canopy resistance
multi-layer soil
roots
sub-soil runoff
surface runoff
prescribed time-varying anthropogenic heat flux
waste heat into urban canopy
buildings & metabolism
traffic & industry

Figure A.8: CM is a multi-layer 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).

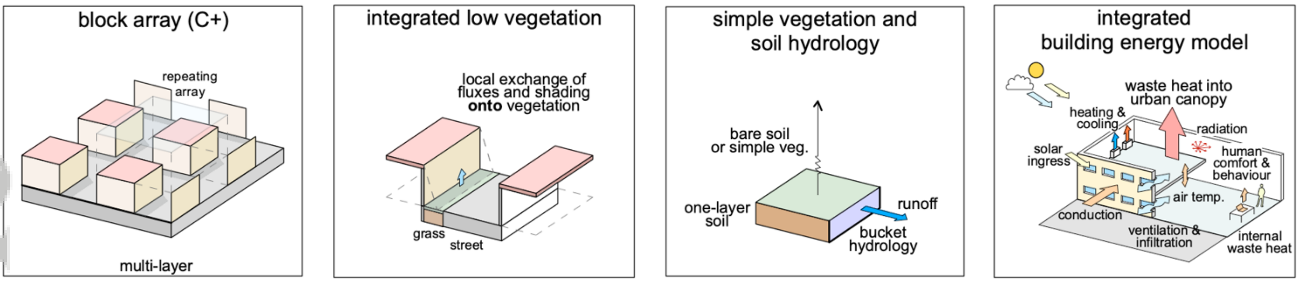

block array (C+)
repeating array
multi-layer
integrated low vegetation
local exchange of fluxes and shading **onto** vegetation
grass
street
simple vegetation and soil hydrology
bare soil or simple veg.
one-layer soil
runoff
bucket hydrology
integrated building energy model
waste heat into urban canopy
heating & cooling
radiation
solar ingress
human comfort & behaviour
air temp.
conduction
ventilation & infiltration
internal waste heat

Figure A.9: CM-BEM is CM (Fig. A.8) 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).

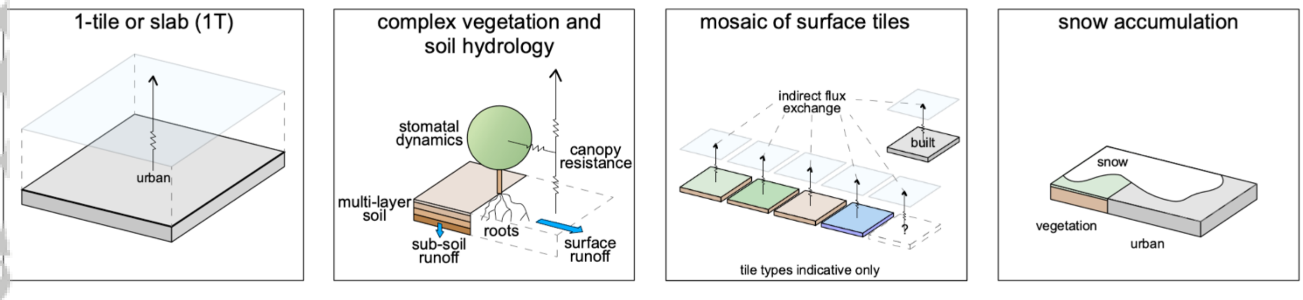


Figure A.10 JULES_1T is a one-tile urban scheme (Best, 2005) within the Joint UK Land Environment Simulator (JULES) (Best et al., 2011). JULES is a community land surface model forming part of the UK Met Office's Unified Model (UM). JULES has nine non-urban surface tiles (five vegetation types, inland 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 longwave radiation only.

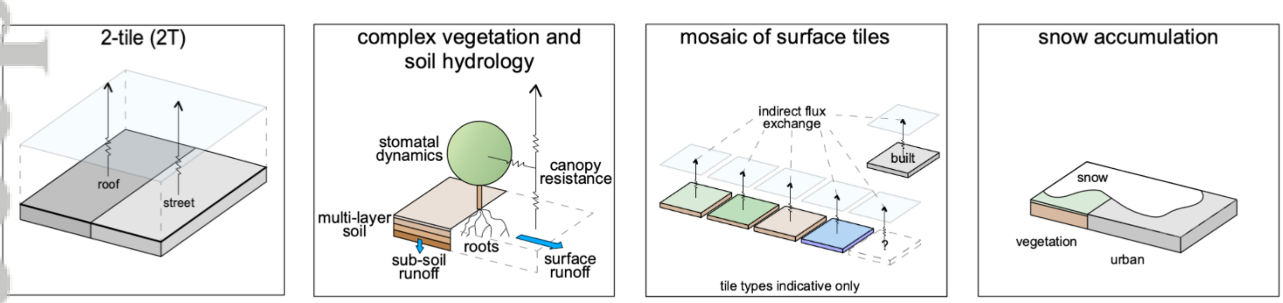


Figure A.11 JULES_2T is a two-urban tile (roof, canyon) (Best et al., 2006) version of JULES_1T (Fig. A.10).

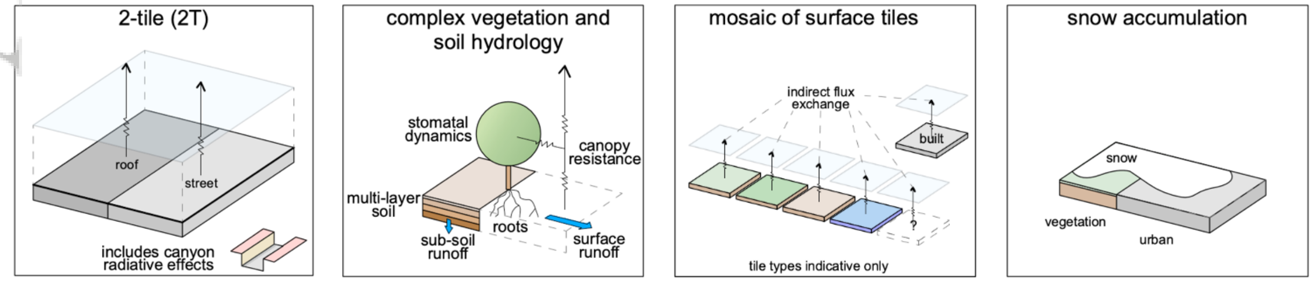


Figure A.12 JULES_MORUSES, has the same two urban tiles (roof, canyon) as JULES_2T (Fig. A.11) but uses MORUSES (Porson et al., 2010). MORUSES has only two facets (roof and canyon) but includes a parameterisation for canyon radiative effects which

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.

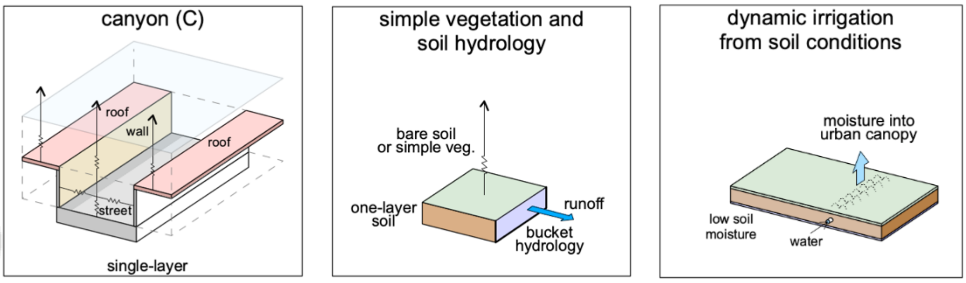


Figure A.13: K-UCMv1 (Klimaat Urban Canopy Model v1) solves the local surface energy balance based on land cover input and regional meteorological forcing. To calculate conduction the urban facets (ground, roof, walls) have 10 layers. Effects of low vegetation are accounted for in the surface energy balance without shadowing effects. Vegetation is perfectly watered and non-vegetation surfaces store no water.

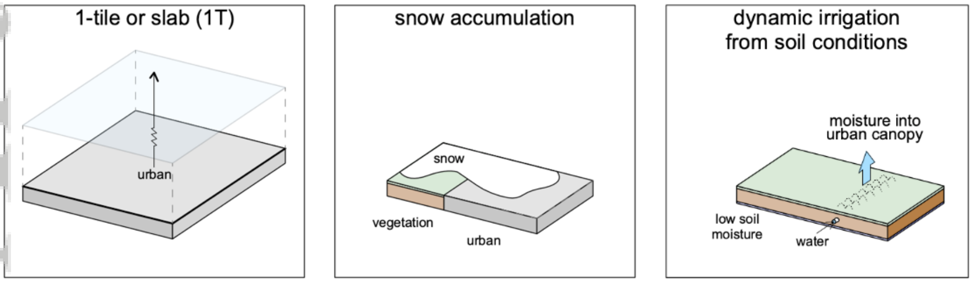
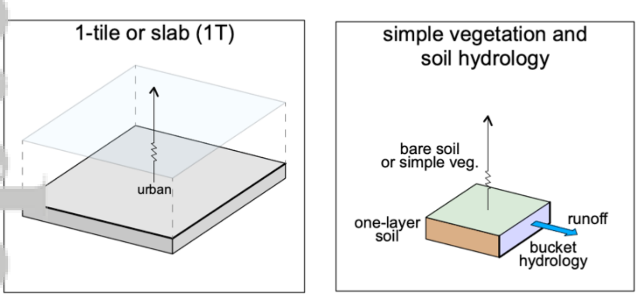


Figure A.14: Lodz-SUEB (Łódź SUrface Energy Balance model) (Fortuniak, 2003) is a bulk scheme that aggregates urban and natural surfaces parameters based on surface fraction. The ground heat flux assumes a 12-layer slab. This is a single snow layer and any surface liquid water or snow, if present, assumed 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.



igure A.15: 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.

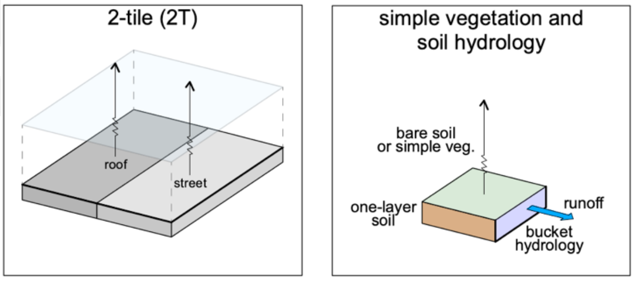


Figure A.16: Manabe_2T only differs from Manabe_1T (Fig. A.15) in that its uses a two-tile urban scheme (roof, canyon) (Best et al., 2006).

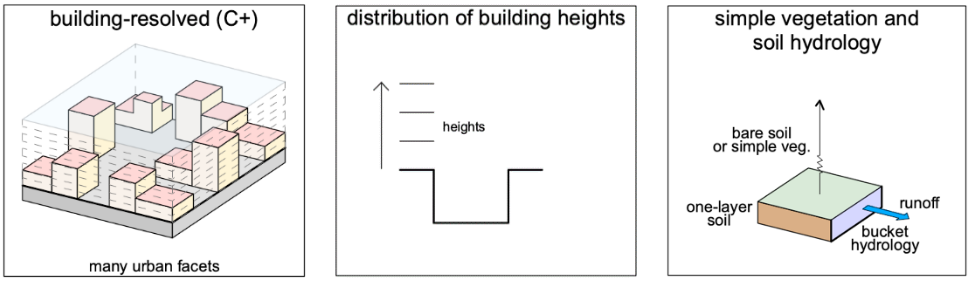


Figure A.17: MUSE (Microscale Urban Surface Energy) (Lee and Lee, 2020) is a building-resolving microscale urban surface model for real urban meteorological and environmental applications. It represents urban buildings on a 3-dimensional Cartesian grid and solves urban physical processes of shortwave and longwave 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.

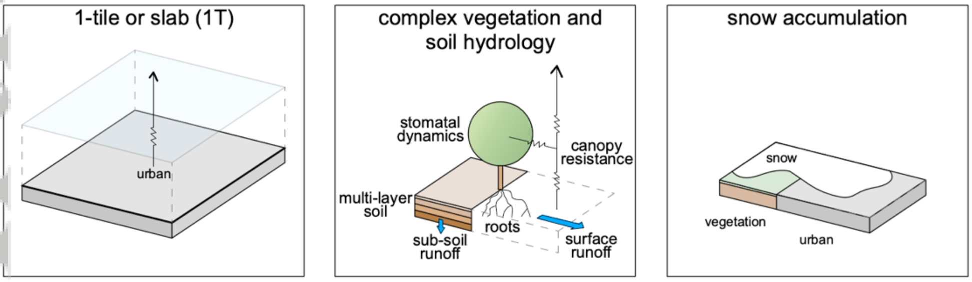


Figure A.18: NOAH-SLAB uses a bulk one-tile urban scheme (Liu et al., 2006) with the Noah land surface model (Noah-LSM) (Chen and Dudhia, 2001). Separate urban and non-urban energy and water balances are simulated then weighted by surface fraction. Although Noah-LSM has up to 27 land-use tiles (including 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.

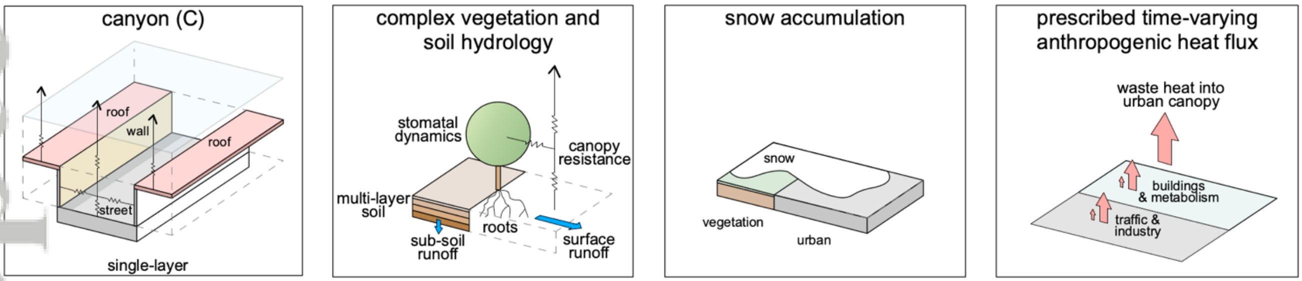


Figure A.19: NOAH-SLUCM uses Noah-LSM as in NOAH-SLAB (Fig. A.18) but uses the Single-Layer Urban Canopy Model (Chen et al., 2011; Kusaka et al., 2001) rather than the urban slab scheme. SLUCM separates the urban tile into three facets (roof, road, wall) using a 2D-canyon approach but without street orientation or varying building heights. A diurnally varying anthropogenic heat flux is prescribed.

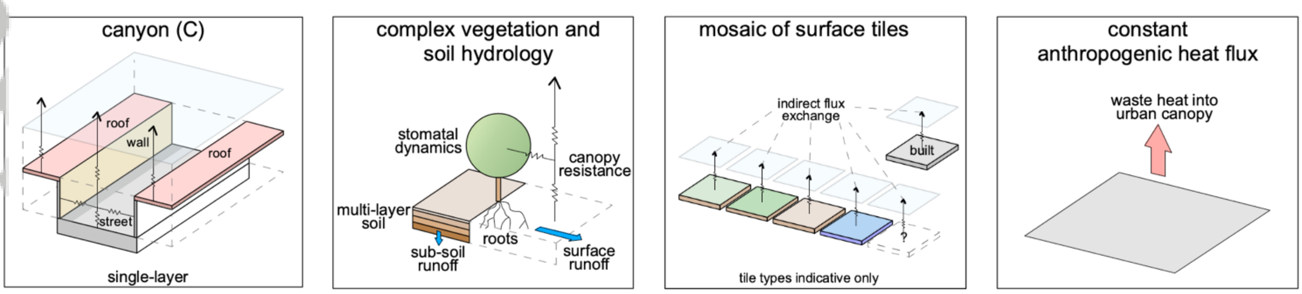


Figure A.20: SNUUCM (Seoul National University Urban Canopy Model), a single-layer model, parameterises shortwave and longwave radiation absorption and reflection, energy and moisture turbulent exchanges between surfaces and adjacent air, and conductive heat transfer through sublayers (Ryu et al., 2011). 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).

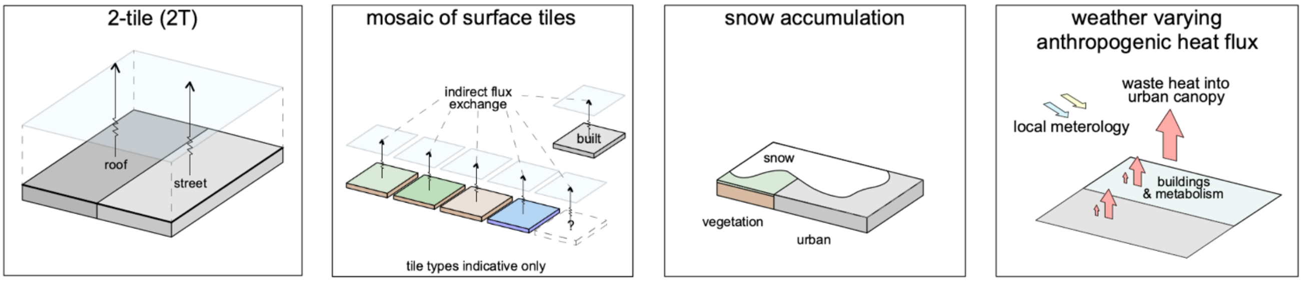


Figure A.21: SUEWS (Surface Urban Energy and Water Balance Scheme) (Järvi et al., 2011; Ward et al., 2016) has two impervious (buildings, paved areas) and five pervious (evergreen trees and shrubs, deciduous trees and shrubs, grass, bare soil and water) surface types, underneath which is a single vertical 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.

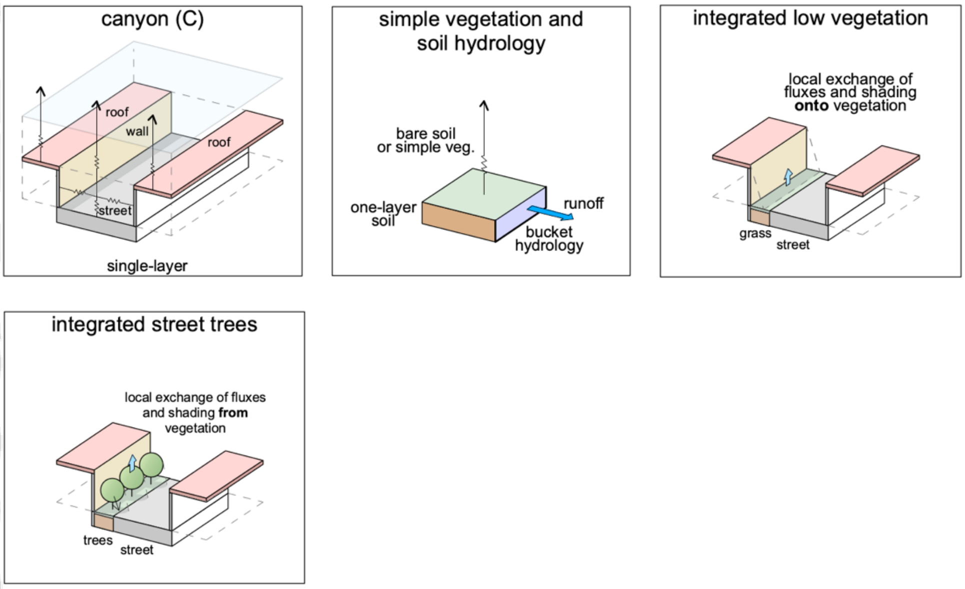

canyon (C)
roof
wall
roof
street
single-layer
simple vegetation and soil hydrology
bare soil or simple veg.
one-layer soil
runoff
bucket hydrology
integrated low vegetation
local exchange of fluxes and shading **onto** vegetation
grass
street
integrated street trees
local exchange of fluxes and shading **from** vegetation
trees
street

Figure A.22: 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 idealized 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.

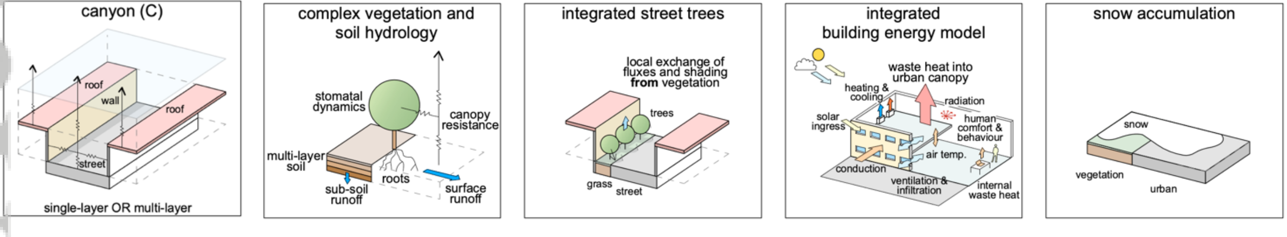


Figure A.23: TEB-CNRM uses the multi-layer version (Hamdi and Masson, 2008; Schoetter et al., 2017) of the urban canopy model Town Energy Balance (Masson, 2000) and part of Météo-France's SURFEX Land Surface Model (Masson et al., 2013; Le Moigne et al., 2018). Here TEB buildings (roofs, walls), roads 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).

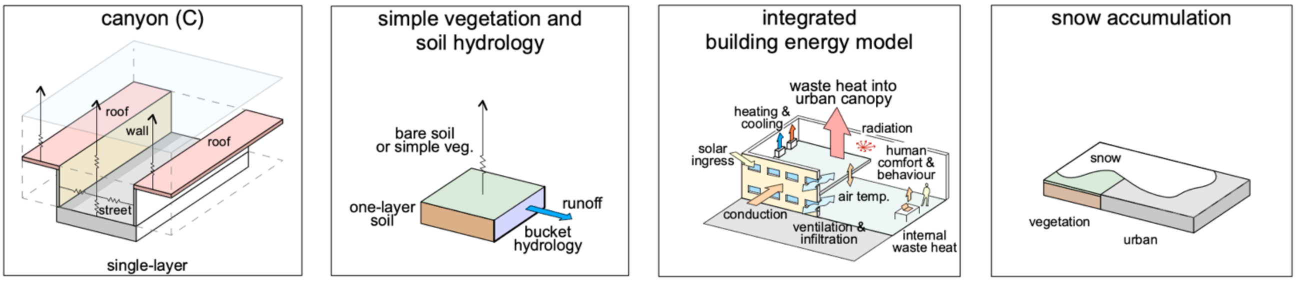


Figure A.24: TEB-READING uses the offline single-layer Town Energy Balance model (Masson, 2000) software (Meyer et al., 2020a) version 4.1.0 (Masson et al., 2021). TEB 4.1.0 is similar to the single layer TEB used in SURFEX version 8.1 (Le Moigne et al., 2018) but with a simple vegetation scheme and time-constant Bowen ratio, albedo, roughness, soil temperature and water availability (Meyer et al., 2020a, b). 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.

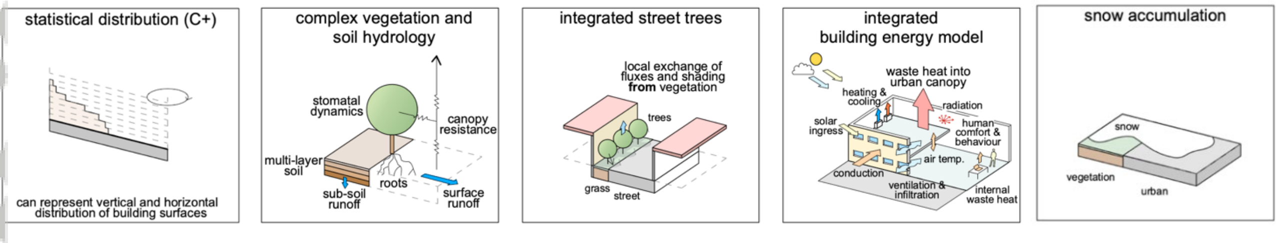

statistical distribution (C+)
can represent vertical and horizontal distribution of building surfaces
complex vegetation and soil hydrology
stomatal dynamics
canopy resistance
multi-layer soil
roots
sub-soil runoff
surface runoff
integrated street trees
local exchange of fluxes and shading from vegetation
trees
grass
street
integrated building energy model
waste heat into urban canopy
heating & cooling
radiation
solar ingress
human comfort & behaviour
air temp.
conduction
ventilation & infiltration
internal waste heat
snow accumulation
snow
vegetation
urban

Fig. A.25: TEB-SPARTCS is builds on TEB-CNRM (Fig. A.23) by incorporating the SPARTACUS-Urban radiative transfer within the urban canopy using a discrete ordinate method (Hogan, 2019a, b), 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.

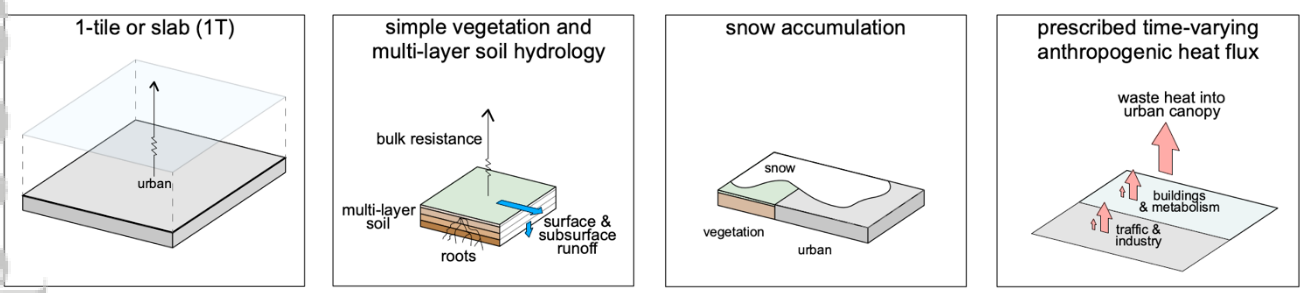

1-tile or slab (1T)
urban
simple vegetation and multi-layer soil hydrology
bulk resistance
multi-layer soil
roots
surface & subsurface runoff
snow accumulation
snow
vegetation
urban
prescribed time-varying anthropogenic heat flux
waste heat into urban canopy
buildings & metabolism
traffic & industry

Fig. A.26: 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 (eg. 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.

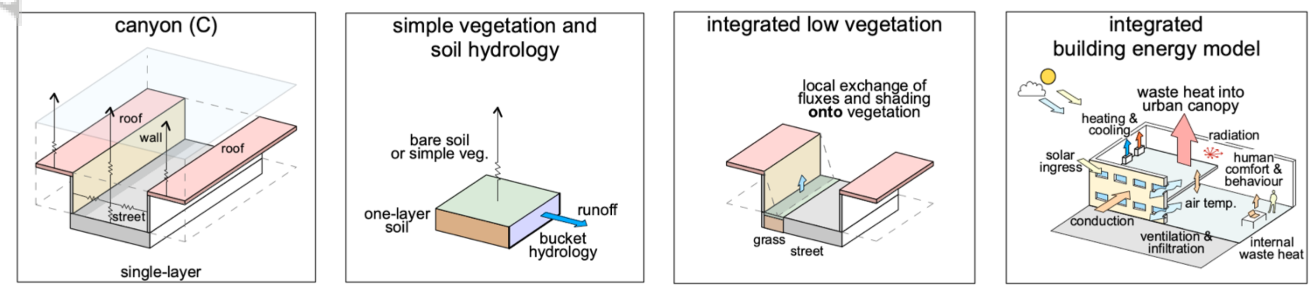

canyon (C)
roof
wall
roof
street
single-layer
simple vegetation and soil hydrology
bare soil or simple veg.
one-layer soil
runoff
bucket hydrology
integrated low vegetation
local exchange of fluxes and shading onto vegetation
grass
street
integrated building energy model
waste heat into urban canopy
heating & cooling
radiation
solar ingress
human comfort & behaviour
air temp.
conduction
ventilation & infiltration
internal waste heat

Fig. A.27: 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 when soil moisture approaches wilting point.

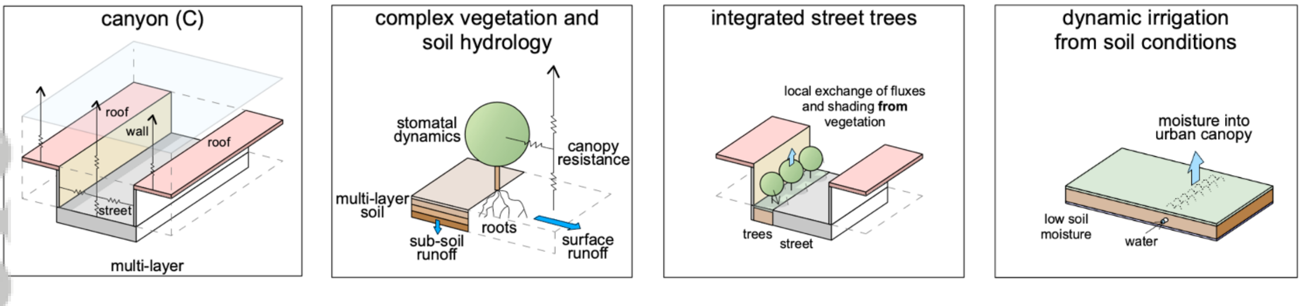


Figure A.28 UT&C: (Urban Tethys-Chloris)(Meili et al., 2020) combines an urban canyon approach with ecohydrological principles of Tethys-Chloris (Fatichi et al., 2012) . Vegetation can occur on roofs and within the canyon (i.e. ground vegetation and/or street trees). Separate soil columns occur below 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.

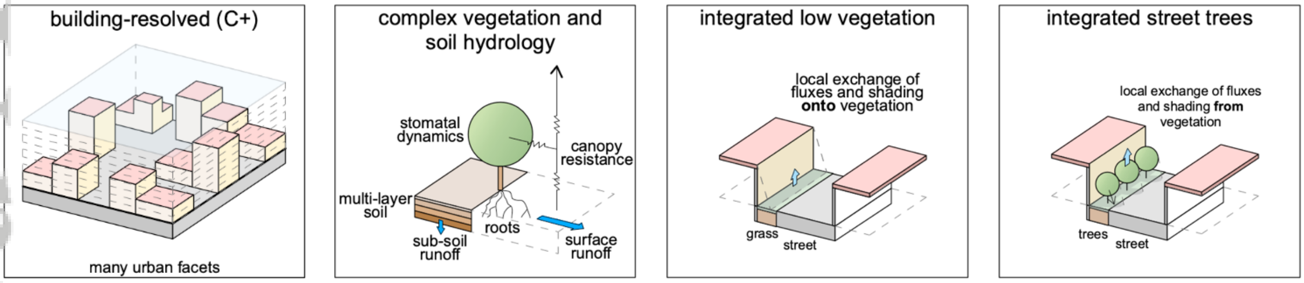


Figure A.29: VTUF-3D (Vegetated Temperatures of Urban Facets in 3D)(Nice et al., 2018) resolves energy transfers in three dimensions by combining TUF-3D (Krayenhoff and Voogt, 2007) with the MAESPA tree model (Duursma and Medlyn, 2012). The vegetation shading and physiological processes are directly integrated with building and urban effects, allowing the role vegetation and water to be assessed in human thermal comfort in urban areas.

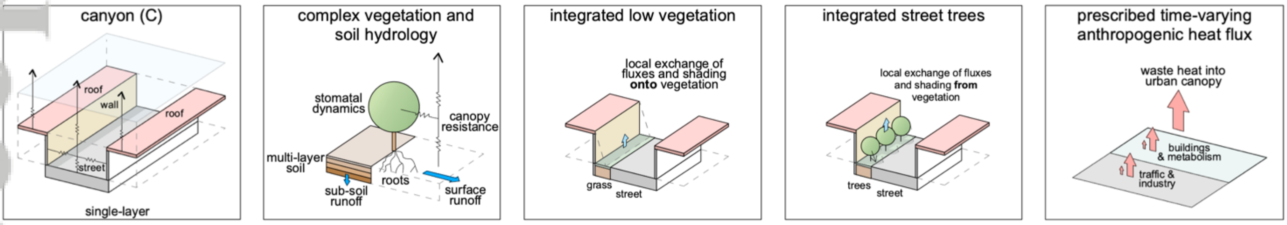

canyon (C)
roof
wall
roof
street
single-layer
complex vegetation and soil hydrology
stomatal dynamics
canopy resistance
multi-layer soil
roots
sub-soil runoff
surface runoff
integrated low vegetation
local exchange of fluxes and shading onto vegetation
grass
street
integrated street trees
local exchange of fluxes and shading from vegetation
trees
street
prescribed time-varying anthropogenic heat flux
waste heat into urban canopy
buildings & metabolism
traffic & industry

Figure A.30: 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 2-dimensional single canyon and a new single tree canopy concept.

10. References

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

Balsamo, G., Beljaars, A., Scipal, K., Viterbo, P., Hurk, B. van den, Hirschi, M., and Betts, A. K.: 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, https://doi.org/10.1175/2008JHM1068.1, 2009.

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

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

Best, M. J.: Progress towards better weather forecasts for city dwellers: from short range to climate change, Theor. Appl. Climatol., 84, 47–55, https://doi.org/10.1007/s00704-005-0143-2, 2006.

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

Best, M. J. and Grimmond, C. S. B.: Key conclusions of the first international urban land surface model comparison project, Bull. Amer. Meteor. Soc., https://doi.org/10.1175/BAMS-D-14-00122.1, 2015.

Best, M. J. and Grimmond, C. S. B.: Investigation of the impact of anthropogenic heat flux within an urban land surface model and PILPS-urban, Theor Appl Climatol, 126, 51–60, https://doi.org/10.1007/s00704-015-1554-3, 2016a.

Best, M. J. and Grimmond, C. S. B.: Modeling the Partitioning of Turbulent Fluxes at Urban Sites with Varying Vegetation Cover, J. Hydrometeor., 17, 2537–2553, https://doi.org/10.1175/JHM-D-15-0126.1, 2016b.

Best, M. J., Grimmond, C. S. B., and Villani, M. G.: Evaluation of the Urban Tile in MOSES using Surface Energy Balance Observations, Boundary-Layer Meteorol, 118, 503–525, https://doi.org/10.1007/s10546-005-9025-5, 2006.

Best, M. J., Pryor, M., Clark, D. B., Rooney, G. G., Essery, R. . L. H., Ménard, C. B., Edwards, J. M., Hendry, M. A., Porson, A., Gedney, N., Mercado, L. M., Sitch, S., Blyth, E., Boucher, O., Cox, P. M., Grimmond, C. S. B., and Harding, R. J.: The Joint UK Land Environment Simulator (JULES), model description – Part 1: Energy and water fluxes, Geosci. Model Dev., 4, 677–699, https://doi.org/10.5194/gmd-4-677-2011, 2011.

Best, M. J., Abramowitz, G., Johnson, H. R., Pitman, A. J., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P. A., Dong, J., Ek, M., Guo, Z., Haverd, V., Hurk, B. J. J. van den, Nearing, G. S., Pak, B., Peters-Lidard, C., Santanello, J. A., Stevens, L., and Vuichard, N.: The Plumbing of Land Surface Models: Benchmarking Model Performance, Journal of Hydrometeorology, 16, 1425–1442, https://doi.org/10.1175/JHM-D-14-0158.1, 2015.

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

Boussetta, S., Balsamo, G., Beljaars, A., Panareda, A.-A., Calvet, J.-C., Jacobs, C., Hurk, B. van den, Viterbo, P., Lafont, S., Dutra, E., Jarlan, L., Balzarolo, M., Papale, D., and Werf, G. van der: Natural land carbon dioxide exchanges in the ECMWF integrated forecasting system: Implementation and offline validation, Journal of Geophysical Research: Atmospheres, 118, 5923–5946, https://doi.org/10.1002/jgrd.50488, 2013.

Bowling, L. and Polcher, J.: The ALMA data exchange convention, https://www.lmd.jussieu.fr/~polcher/ALMA/, 2001.

Bowling, L. C., Lettenmaier, D. P., Nijssen, B., Graham, L. P., Clark, D. B., El Maayar, M., Essery, R., Goers, S., Gusev, Y. M., Habets, F., van den Hurk, B., Jin, J., Kahan, D., Lohmann, D., Ma, X., Mahanama, S., Mocko, D., Nasonova, O., Niu, G.-Y., Samuelsson, P., Shmakin, A. B., Takata, K., Verseghy, D., Viterbo, P., Xia, Y., Xue, Y., and Yang, Z.-L.: 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, https://doi.org/10.1016/S0921-8181(03)00003-1, 2003.

Broadbent, A. M., Coutts, A. M., Nice, K. A., Demuzere, M., Krayenhoff, E. S., Tapper, N. J., and Wouters, H.: 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, https://doi.org/10.5194/gmd-12-785-2019, 2019.

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

Chen, F. and Dudhia, J.: 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, https://doi.org/10.1175/1520-0493(2001)129<0569:CAALSH>2.0.CO;2, 2001.

Chen, F., Kusaka, H., Bornstein, R., Ching, J., Grimmond, C. S. B., Grossman-Clarke, S., Loridan, T., Manning, K. W., Martilli, A., Miao, S., Sailor, D., Salamanca, F., Taha, H., Tewari, M., Wang, X., Wyszogrodzki, A. A., and Zhang, C.: The integrated WRF/urban modelling system: development, evaluation, and applications to urban environmental problems, Int. J. Climatol., 31, 273–288, https://doi.org/10.1002/joc.2158, 2011.

Chow, W.: 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, https://doi.org/10.6073/PASTA/FED17D67583EDA16C439216CA40B0669, 2017.

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

Christen, A., Coops, N. C., Crawford, B. R., Kellett, R., Liss, K. N., Olchovski, I., Tooke, T. R., van der Laan, M., and Voogt, J. A.: Validation of modeled carbon-dioxide emissions from an urban neighborhood with direct eddy-covariance measurements, Atmospheric Environment, 45, 6057–6069, https://doi.org/10.1016/j.atmosenv.2011.07.040, 2011.

Coutts, A. M.: 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--Monash University, 2006.

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

Coutts, A. M., Beringer, J., and Tapper, N. J.: Impact of Increasing Urban Density on Local Climate: Spatial and Temporal Variations in the Surface Energy Balance in Melbourne, Australia, J. Appl. Meteor. Climatol., 46, 477–493, https://doi.org/10.1175/JAM2462.1, 2007b.

Crawford, B. and Christen, A.: Spatial source attribution of measured urban eddy covariance CO2 fluxes, Theor Appl Climatol, 119, 733–755, https://doi.org/10.1007/s00704-014-1124-0, 2015.

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

Cucchi, M., Weedon, G. P., Amici, A., Bellouin, N., Lange, S., Schmied, H. M., Hersbach, H., and Buontempo, C.: WFDE5: bias adjusted ERA5 reanalysis data for impact studies, Earth System Science Data Discussions, 1–32, https://doi.org/10.5194/essd-2020-28, 2020.

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

Demuzere, M., Kittner, J., Martilli, A., Mills, G., Moede, C., Stewart, I. D., van Vliet, J., and Bechtel, B.: A global map of Local Climate Zones to support earth system modelling and urban scale environmental science, Earth System Science Data Discussions, 1–57, https://doi.org/10.5194/essd-2022-92, 2022.

Duursma, R. A. and Medlyn, B. E.: 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] × drought interactions, Geoscientific Model Development, 5, 919–940, https://doi.org/10.5194/gmd-5-919-2012, 2012.

ECMWF: IFS Documentation CY47R1 - Part IV: Physical Processes, https://doi.org/10.21957/CPMKQVHJA, 2020.

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., and Taylor, K. E.: Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization, Geoscientific Model Development, 9, 1937–1958, https://doi.org/10.5194/gmd-9-1937-2016, 2016.

Fatichi, S., Ivanov, V. Y., and Caporali, E.: 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, https://doi.org/10.1029/2011MS000086, 2012.

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

Fortuniak, K.: 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, 2003.

Fortuniak, K., Kłysik, K., and Siedlecki, M.: 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, 64–67, 2006.

Fortuniak, K., Pawlak, W., and Siedlecki, M.: Integral Turbulence Statistics Over a Central European City Centre, Boundary Layer Meteorology; Dordrecht, 146, 257–276, https://doi.org/10.1007/s10546-012-9762-1, 2013.

Garbero, V., Milelli, M., Bucchignani, E., Mercogliano, P., Varentsov, M., Rozinkina, I., Rivin, G., Blinov, D., Wouters, H., Schulz, J.-P., Schättler, U., Bassani, F., Demuzere, M., and Repola, F.: Evaluating the Urban Canopy Scheme TERRA_URB in the COSMO Model for Selected European Cities, Atmosphere, 12, 237, https://doi.org/10.3390/atmos12020237, 2021.

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

Goret, M., Masson, V., Schoetter, R., and Moine, M.-P.: 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, https://doi.org/10.1016/j.aeaoa.2019.100042, 2019.

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

Grimmond, C. S. B., Best, M., Barlow, J., Arnfield, A. J., Baik, J.-J., Baklanov, A., Belcher, S., Bruse, M., Calmet, I., Chen, F., Clark, P., Dandou, A., Erell, E., Fortuniak, K., Hamdi, R., Kanda, M., Kawai, T., Kondo, H., Krayenhoff, S., Lee, S. H., Limor, S.-B., Martilli, A., Masson, V., Miao, S., Mills, G., Moriwaki, R., Oleson, K., Porson, A., Sievers, U., Tombrou, M., Voogt, J., and Williamson, T.: Urban Surface Energy Balance Models: Model Characteristics and Methodology for a Comparison Study, in: Meteorological and Air Quality Models for Urban Areas, edited by: Baklanov, A., Sue, G., Alexander, M., and Athanassiadou, M., Springer Berlin Heidelberg, 97–123, 2009.

Grimmond, C. S. B., Blackett, M., Best, M. J., Barlow, J., Baik, J.-J., Belcher, S. E., Bohnenstengel, S. I., Calmet, I., Chen, F., Dandou, A., Fortuniak, K., Gouvea, M. L., Hamdi, R., Hendry, M., Kawai, T., Kawamoto, Y., Kondo, H., Krayenhoff, E. S., Lee, S.-H., and Loridan, T.: The International Urban Energy Balance Models Comparison Project: First Results from Phase 1, Journal of Applied Meteorology & Climatology, 49, 1268–1292, https://doi.org/10.1175/2010JAMC2354.1, 2010.

Grimmond, C. S. B., Blackett, M., Best, M. J., Baik, J.-J., Belcher, S. E., Beringer, J., Bohnenstengel, S. I., Calmet, I., Chen, F., Coutts, A., Dandou, A., Fortuniak, K., Gouvea, M. L., Hamdi, R., Hendry, M., Kanda, M., Kawai, T., Kawamoto, Y., Kondo, H., Krayenhoff, E. S., Lee, S.-H., Loridan, T., Martilli, A., Masson, V., Miao, S., Oleson, K., Ooka, R., Pigeon, G., Porson, A., Ryu, Y.-H., Salamanca, F., Steeneveld, G. j., Tombrou, M., Voogt, J. A., Young, D. T., and Zhang, N.: Initial results from Phase 2 of the international urban energy balance model comparison, International Journal of Climatology, 31, 244–272, https://doi.org/10.1002/joc.2227, 2011.

Hamdi, R. and Masson, V.: 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, https://doi.org/10.1175/2008JAMC1865.1, 2008.

Haughton, N., Abramowitz, G., Pitman, A. J., Or, D., Best, M. J., Johnson, H. R., Balsamo, G., Boone, A., Cuntz, M., Decharme, B., Dirmeyer, P. A., Dong, J., Ek, M., Guo, Z., Haverd, V., Hurk, B. J. J. van den, Nearing, G. S., Pak, B., Santanello, J. A., Stevens, L. E., and Vuichard, N.: The Plumbing of Land Surface Models: Is Poor Performance a Result of Methodology or Data Quality?, Journal of Hydrometeorology, 17, 1705–1723, https://doi.org/10.1175/JHM-D-15-0171.1, 2016.

Haughton, N., Abramowitz, G., and Pitman, A. J.: On the Predictability of Land Surface Fluxes from Meteorological Variables, Geosci. Model Dev. Discuss., 2017, 1–27, https://doi.org/10.5194/gmd-2017-153, 2017.

Henderson-Sellers, A., Pitman, A. J., Love, P. K., Irannejad, P., and Chen, T. H.: The Project for Intercomparison of Land Surface Parameterization Schemes (PILPS): Phases 2 and 3*, Bulletin of the American Meteorological Society, 76, 489–504, https://doi.org/10.1175/1520-0477(1995)076<0489:TPFIOL>2.0.CO;2, 1995.

Henderson-Sellers, A., McGuffie, K., and Pitman, A. J.: The Project for Intercomparison of Land-surface Parametrization Schemes (PILPS): 1992 to 1995, Climate Dynamics, 12, 849–859, https://doi.org/10.1007/s003820050147, 1996.

Hengl, T.: 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, 2018a.

Hengl, T.: 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, 2018b.

Hengl, T.: Soil bulk density (fine earth) 10 x 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, 2018c.

Hersbach, H., Bell, B., Berrisford, P., Biavati, G., Horányi, A., Muñoz Sabater, J., Nicolas, J., Peubey, C., Radu, R., Rozum, I., and others: ERA5 hourly data on single levels from 1979 to present, 2018.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., Simmons, A., Soci, C., Abdalla, S., Abellan, X., Balsamo, G., Bechtold, P., Biavati, G., Bidlot, J., Bonavita, M., Chiara, G. D., Dahlgren, P., Dee, D., Diamantakis, M., Dragani, R., Flemming, J., Forbes, R., Fuentes, M., Geer, A., Haimberger, L., Healy, S., Hogan, R. J., Hólm, E., Janisková, M., Keeley, S., Laloyaux, P., Lopez, P., Lupu, C., Radnoti, G., Rosnay, P. de, Rozum, I., Vamborg, F., Villaume, S., and Thépaut, J.-N.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, https://doi.org/10.1002/qj.3803, 2020.

Hirano, T., Sugawara, H., Murayama, S., and Kondo, H.: Diurnal Variation of CO_2 Flux in an Urban Area of Tokyo, Sola, 11, 100–103, https://doi.org/10.2151/sola.2015-024, 2015.

Hogan, R. J.: An Exponential Model of Urban Geometry for Use in Radiative Transfer Applications, Boundary-Layer Meteorol, 170, 357–372, https://doi.org/10.1007/s10546-018-0409-8, 2019a.

Hogan, R. J.: Flexible Treatment of Radiative Transfer in Complex Urban Canopies for Use in Weather and Climate Models, Boundary-Layer Meteorol, 173, 53–78, https://doi.org/10.1007/s10546-019-00457-0, 2019b.

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

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

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

Ishidoya, S., Sugawara, H., Terao, Y., Kaneyasu, N., Aoki, N., Tsuboi, K., and Kondo, H.: 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, https://doi.org/10.5194/acp-20-5293-2020, 2020.

Jackson, E. K., Roberts, W., Nelsen, B., Williams, G. P., Nelson, E. J., and Ames, D. P.: 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, https://doi.org/10.1016/j.envsoft.2019.05.001, 2019.

Jackson, T. L., Feddema, J. J., Oleson, K. W., Bonan, G. B., and Bauer, J. T.: Parameterization of Urban Characteristics for Global Climate Modeling, Annals of the Association of American Geographers, 100, 848–865, https://doi.org/10.1080/00045608.2010.497328, 2010.

Järvi, L., Grimmond, C. S. B., and Christen, A.: The Surface Urban Energy and Water Balance Scheme (SUEWS): Evaluation in Los Angeles and Vancouver, Journal of Hydrology, 411, 219–237, https://doi.org/10.1016/j.jhydrol.2011.10.001, 2011.

Järvi, L., Rannik, Ü., Kokkonen, T. V., Kurppa, M., Karppinen, A., Kouznetsov, R. D., Rantala, P., Vesala, T., and Wood, C. R.: Uncertainty of eddy covariance flux measurements over an urban area based on two towers, Atmospheric Measurement Techniques, 11, 5421–5438, https://doi.org/10.5194/amt-11-5421-2018, 2018.

Jo, S., Hong, J.-W., and Hong, J.: The observational flux measurement data of suburban and low-residential areas in Korea (in preparation), n.d.

Kanda, M., Kawai, T., Kanega, M., Moriwaki, R., Narita, K., and Hagishima, A.: A Simple Energy Balance Model for Regular Building Arrays, Boundary-Layer Meteorol, 116, 423–443, https://doi.org/10.1007/s10546-004-7956-x, 2005.

Karsisto, P., Fortelius, C., Demuzere, M., Grimmond, C. S. B., W., O. K., Kouznetsov, R., Masson, V., and Järvi, L.: Seasonal surface urban energy balance and wintertime stability simulated using three land-surface models in the high-latitude city Helsinki, Q.J.R. Meteorol. Soc., 142, 401–417, https://doi.org/10.1002/qj.2659, 2016.

Kikegawa, Y., Genchi, Y., Yoshikado, H., and Kondo, H.: 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, https://doi.org/10.1016/S0306-2619(03)00009-6, 2003.

Kikegawa, Y., Genchi, Y., Kondo, H., and Hanaki, K.: 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, https://doi.org/10.1016/j.apenergy.2005.06.001, 2006.

Kikegawa, Y., Tanaka, A., Ohashi, Y., Ihara, T., and Shigeta, Y.: Observed and simulated sensitivities of summertime urban surface air temperatures to anthropogenic heat in downtown areas of two Japanese Major Cities, Tokyo and Osaka, Theor Appl Climatol, 117, 175–193, https://doi.org/10.1007/s00704-013-0996-8, 2014.

Kondo, H. and Liu, F.-H.: 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, https://doi.org/10.11298/taiki1995.33.3_179, 1998.

Kondo, H., Genchi, Y., Kikegawa, Y., Ohashi, Y., Yoshikado, H., and Komiyama, H.: 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, https://doi.org/10.1007/s10546-005-0905-5, 2005.

Kondo, H., Inagaki, A., and Kanda, M.: 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, https://doi.org/10.1007/s10546-015-0019-7, 2015.

Koster, R. D., Guo, Z., Yang, R., Dirmeyer, P. A., Mitchell, K., and Puma, M. J.: On the Nature of Soil Moisture in Land Surface Models, Journal of Climate, 22, 4322–4335, https://doi.org/10.1175/2009JCLI2832.1, 2009.

Kotthaus, S. and Grimmond, C. S. B.: Energy exchange in a dense urban environment – Part I: Temporal variability of long-term observations in central London, Urban Climate, 10, Part 2, 261–280, https://doi.org/10.1016/j.uclim.2013.10.002, 2014a.

Kotthaus, S. and Grimmond, C. S. B.: Energy exchange in a dense urban environment – Part II: Impact of spatial heterogeneity of the surface, Urban Climate, 10, Part 2, 281–307, https://doi.org/10.1016/j.uclim.2013.10.001, 2014b.

Kowalczyk, E., Stevens, L., Law, R., Dix, M., Wang, Y., Harman, I., Haynes, K., Srbinovsky, J., Pak, B., and Ziehn, T.: The land surface model component of ACCESS: description and impact on the simulated surface climatology, Australian Meteorological and Oceanographic Journal, 63, 65–82, 2013.

Kowalczyk, E. A., Wang, Y. P., Law, R. M., Davies, H. L., McGregor, J. L., and Abramowitz, G. S.: 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, 2006.

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

Kusaka, H., Kondo, H., Kikegawa, Y., and Kimura, F.: A Simple Single-Layer Urban Canopy Model For Atmospheric Models: Comparison With Multi-Layer And Slab Models, Boundary-Layer Meteorology, 101, 329–358, https://doi.org/10.1023/A:1019207923078, 2001.

Le Moigne, P., C. Albergel, A. Boone, S. Belamari, B. Decharme, M. Dumont, P. Le Moigne, and V. Masson: SURFEX v8.1 Scientific Documentation, 2018.

Lee, D.-I. and Lee, S.-H.: The Microscale Urban Surface Energy (MUSE) Model for Real Urban Application, Atmosphere, 11, 1347, https://doi.org/10.3390/atmos11121347, 2020.

Lee, S.-H.: Further Development of the Vegetated Urban Canopy Model Including a Grass-Covered Surface Parametrization and Photosynthesis Effects, Boundary-Layer Meteorol, 140, 315–342, https://doi.org/10.1007/s10546-011-9603-7, 2011.

Lee, S.-H. and Park, S.-U.: A Vegetated Urban Canopy Model for Meteorological and Environmental Modelling, Boundary-Layer Meteorol, 126, 73–102, https://doi.org/10.1007/s10546-007-9221-6, 2008.

Lee, S.-H., Lee, H., Park, S.-B., Woo, J.-W., Lee, D.-I., and Baik, J.-J.: 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, https://doi.org/10.1002/qj.2847, 2016.

Lemonsu, A., Masson, V., Shashua-Bar, L., Erell, E., and Pearlmutter, D.: Inclusion of vegetation in the Town Energy Balance model for modelling urban green areas, Geosci. Model Dev., 5, 1377–1393, https://doi.org/10.5194/gmd-5-1377-2012, 2012.

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

Lipson, M., Grimmond, S., Best, M., Abramowitz, G., Kauwe, M. D., Tsiringakis, A., Demuzere, M., Ward, H., Coutts, A., and Pitman, A.: Modelling protocol for the Urban-PLUMBER model evaluation project, https://doi.org/10.5281/zenodo.6363850, 2020.

Lipson, M., Grimmond, S., Best, M., Chow, W. T. L., Christen, A., Chrysoulakis, N., Coutts, A., Crawford, B., Earl, S., Evans, J., Fortuniak, K., Heusinkveld, B. G., Hong, J.-W., Hong, J., Järvi, L., Jo, S., Kim, Y.-H., Kotthaus, S., Lee, K., Masson, V., McFadden, J. P., Michels, O., Pawlak, W., Roth, M., Sugawara, H., Tapper, N., Velasco, E., and Ward, H. C.: Harmonized gap-filled datasets from 20 urban flux tower sites, Earth System Science Data, 14, 5157–5178, https://doi.org/10.5194/essd-14-5157-2022, 2022a.

Lipson, M., Grimmond, S., Best, M., Abramowitz, G., Coutts, A., Tapper, N., Baik, J.-J., Beyers, M., Blunn, L., Boussetta, S., bou-Zeid, E., Kauwe, M. G. D., Munck, C. de, Demuzere, M., Fatichi, S., Fortuniak, K., Han, B.-S., Hendry, M., Kikegawa, Y., Kondo, H., Lee, D.-I., Lee, S.-H., Lemonsu, A., Machado, T., Manoli, G., Martilli, A., Masson, V., McNorton, J., Meili, N., Meyer, D., Nice, K. A., Oleson, K. W., Park, S.-B., Roth, M., Schoetter, R., Simon, A., Steeneveld, G.-J., Sun, T., Takane, Y., Thatche, M., Tsiringakis, A., Varentsov, M., Wang, C., and Wang, Z.-H.: Associated results of Phase 1 of the Urban-PLUMBER model evaluation project, https://doi.org/10.5281/zenodo.7388342, 2022b.

Lipson, M., Grimmond, S., Best, M., Chow, W., Christen, A., Chrysoulakis, N., Coutts, A., Crawford, B., Earl, S., Evans, J., Fortuniak, K., Heusinkveld, B. G., Hong, J.-W., Hong, J., Järvi, L., Jo, S., Kim, Y.-H., Kotthaus, S., Lee, K., Masson, V., McFadden, J. P., Michels, O., Pawlak, W., Roth, M., Sugawara, H., Tapper, N., Velasco, E., and Ward, H. C.: Data for "Harmonized gap-filled dataset from 20 urban flux tower sites" for the Urban-PLUMBER project, https://doi.org/10.5281/zenodo.7104984, 2022c.

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

Liu, Y., Chen, F., Warner, T., and Basara, J.: 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, https://doi.org/10.1175/JAM2383.1, 2006.

Manabe, S.: CLIMATE AND THE OCEAN CIRCULATION: I. THE ATMOSPHERIC CIRCULATION AND THE HYDROLOGY OF THE EARTH'S SURFACE, Monthly Weather Review, 97, 739–774, https://doi.org/10.1175/1520-0493(1969)097<0739:CATOC>2.3.CO;2, 1969.

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

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

Martilli, A., Santiago, J. L., and Salamanca, F.: On the representation of urban heterogeneities in mesoscale models, Environ Fluid Mech, 15, 305–328, https://doi.org/10.1007/s10652-013-9321-4, 2015.

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

Masson, V., Gomes, L., Pigeon, G., Liousse, C., Pont, V., Lagouarde, J.-P., Voogt, J., Salmond, J., Oke, T. R., Hidalgo, J., Legain, D., Garrouste, O., Lac, C., Connan, O., Briottet, X., Lachérade, S., and Tulet, P.: The Canopy and Aerosol Particles Interactions in TOulouse Urban Layer (CAPITOUL) experiment, Meteorol Atmos Phys, 102, 135–157, https://doi.org/10.1007/s00703-008-0289-4, 2008.

Masson, V., Le Moigne, P., Martin, E., Faroux, S., Alias, A., Alkama, R., Belamari, S., Barbu, A., Boone, A., Bouyssel, F., Brousseau, P., Brun, E., Calvet, J.-C., Carrer, D., Decharme, B., Delire, C., Donier, S., Essaouini, K., Gibelin, A.-L., Giordani, H., Habets, F., Jidane, M., Kerdraon, G., Kourzeneva, E., Lafaysse, M., Lafont, S., Lebeaupin Brossier, C., Lemonsu, A., Mahfouf, J.-F., Marguinaud, P., Mokhtari, M., Morin, S., Pigeon, G., Salgado, R., Seity, Y., Taillefer, F., Tanguy, G., Tulet, P., Vincendon, B., Vionnet, V., and Voldoire, A.: 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, https://doi.org/10.5194/gmd-6-929-2013, 2013.

Masson, V., Lemonsu, A., Hidalgo, J., and Voogt, J.: Urban Climates and Climate Change, Annual Review of Environment and Resources, 45, 411–444, https://doi.org/10.1146/annurev-environ-012320-083623, 2020.

Masson, V., Lemonsu, A., Pigeon, G., Schoetter, R., de Munck, C., Bueno, B., Faroux, S., Goret, M., Redon, E., Chancibault, K., Stavropulos-Laffaille, X., Leroyer, S., and Meyer, D.: The Town Energy Balance (TEB) model, , https://doi.org/10.5281/zenodo.5104731, 2021.

McNorton, J. R., Arduini, G., Bousserez, N., Agustí-Panareda, A., Balsamo, G., Boussetta, S., Choulga, M., Hadade, I., and Hogan, R. J.: An Urban Scheme for the ECMWF Integrated Forecasting System: Single-Column and Global Offline Application, Journal of Advances in Modeling Earth Systems, 13, e2020MS002375, https://doi.org/10.1029/2020MS002375, 2021.

Meili, N., Manoli, G., Burlando, P., Bou-Zeid, E., Chow, W. T. L., Coutts, A. M., Daly, E., Nice, K. A., Roth, M., Tapper, N. J., Velasco, E., Vivoni, E. R., and Fatichi, S.: An urban ecohydrological model to quantify the effect of vegetation on urban climate and hydrology (UT&amp;C v1.0), Geoscientific Model Development, 13, 335–362, https://doi.org/10.5194/gmd-13-335-2020, 2020.

Meili, N., Manoli, G., Burlando, P., Carmeliet, J., Chow, W. T. L., Coutts, A. M., Roth, M., Velasco, E., Vivoni, E. R., and Fatichi, S.: 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, https://doi.org/10.1016/j.ufug.2020.126970, 2021.

Menard, C. B., Essery, R., Krinner, G., Arduini, G., Bartlett, P., Boone, A., Brutel-Vuilmet, C., Burke, E., Cuntz, M., Dai, Y., Decharme, B., Dutra, E., Fang, X., Fierz, C., Gusev, Y., Hagemann, S., Haverd, V., Kim, H., Lafaysse, M., Marke, T., Nasonova, O., Nitta, T., Niwano, M., Pomeroy, J., Schädler, G., Semenov, V. A., Smirnova, T., Strasser, U., Swenson, S., Turkov, D., Wever, N., and Yuan, H.: Scientific and Human Errors in a Snow Model Intercomparison, Bulletin of the American Meteorological Society, 102, E61–E79, https://doi.org/10.1175/BAMS-D-19-0329.1, 2021.

Menzer, O. and McFadden, J. P.: 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, https://doi.org/10.1016/j.atmosenv.2017.09.049, 2017.

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

Meyer, D., Schoetter, R., Riechert, M., Verrelle, A., Tewari, M., Dudhia, J., Masson, V., Reeuwijk, M. van, and Grimmond, S.: 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, https://doi.org/10.1029/2019MS001961, 2020a.

Meyer, D., Schoetter, R., Masson, V., and Grimmond, S.: Enhanced software and platform for the Town Energy Balance (TEB) model, JOSS, 5, 2008, https://doi.org/10.21105/joss.02008, 2020b.

Nazarian, N., Lipson, M., and Norford, L. K.: Chapter 4 - Multiscale modeling techniques to document urban climate change, in: Urban Climate Change and Heat Islands, edited by: Paolini, R. and Santamouris, M., Elsevier, 123–164, https://doi.org/10.1016/B978-0-12-818977-1.00004-1, 2023.

Nice, K. A., Coutts, A. M., and Tapper, N. J.: 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, https://doi.org/10.1016/j.uclim.2017.12.008, 2018.

Nordbo, A., Järvi, L., Haapanala, S., Moilanen, J., and Vesala, T.: Intra-City Variation in Urban Morphology and Turbulence Structure in Helsinki, Finland, Boundary-Layer Meteorol, 146, 469–496, https://doi.org/10.1007/s10546-012-9773-y, 2013.

Oleson, K. W. and Feddema, J.: Parameterization and Surface Data Improvements and New Capabilities for the Community Land Model Urban (CLMU), Journal of Advances in Modeling Earth Systems, 12, https://doi.org/10.1029/2018MS001586, 2020.

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

Oleson, K. W., Anderson, G. B., Jones, B., McGinnis, S. A., and Sanderson, B.: 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, https://doi.org/10.1007/s10584-015-1504-1, 2018.

Pawlak, W., Fortuniak, K., and Siedlecki, M.: 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, https://doi.org/10.1002/joc.2247, 2011.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É.: Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research, 2011.

Perkins, S. E., Pitman, A. J., Holbrook, N. J., and McAneney, J.: Evaluation of the AR4 Climate Models' Simulated Daily Maximum Temperature, Minimum Temperature, and Precipitation over Australia Using Probability Density Functions, J. Climate, 20, 4356–4376, https://doi.org/10.1175/JCLI4253.1, 2007.

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

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

Porson, A., Clark, P. A., Harman, I. N., Best, M. J., and Belcher, S. E.: 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, https://doi.org/10.1002/qj.668, 2010.

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

Richardson, A. D., Hollinger, D. Y., Burba, G. G., Davis, K. J., Flanagan, L. B., Katul, G. G., William Munger, J., Ricciuto, D. M., Stoy, P. C., Suyker, A. E., Verma, S. B., and Wofsy, S. C.: A multi-site analysis of random error in tower-based measurements of carbon and energy fluxes, Agricultural and Forest Meteorology, 136, 1–18, https://doi.org/10.1016/j.agrformet.2006.01.007, 2006.

Rockel, B., Will, A., and Hense, A.: The Regional Climate Model COSMO-CLM (CCLM), Meteorologische Zeitschrift, 347–348, https://doi.org/10.1127/0941-2948/2008/0309, 2008.

Roth, M., Jansson, C., and Velasco, E.: Multi-year energy balance and carbon dioxide fluxes over a residential neighbourhood in a tropical city, Int. J. Climatol., 37, 2679–2698, https://doi.org/10.1002/joc.4873, 2017.

Ryu, Y.-H., Baik, J.-J., and Lee, S.-H.: A New Single-Layer Urban Canopy Model for Use in Mesoscale Atmospheric Models, Journal of Applied Meteorology and Climatology, 50, 1773–1794, https://doi.org/10.1175/2011JAMC2665.1, 2011.

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

Sailor, D. J.: A review of methods for estimating anthropogenic heat and moisture emissions in the urban environment, Int. J. Climatol., 31, 189–199, https://doi.org/10.1002/joc.2106, 2011.

Salamanca, F., Krayenhoff, E. S., and Martilli, A.: On the Derivation of Material Thermal Properties Representative of Heterogeneous Urban Neighborhoods, Journal of Applied Meteorology & Climatology, 48, 1725–1732, https://doi.org/10.1175/2009JAMC2176.1, 2009.

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

Schoetter, R., Masson, V., Bourgeois, A., Pellegrino, M., and Lévy, J.-P.: Parametrisation of the variety of human behaviour related to building energy consumption in the Town Energy Balance (SURFEX-TEB v. 8.2), Geosci. Model Dev., 10, 2801–2831, https://doi.org/10.5194/gmd-10-2801-2017, 2017.

Schulz, J.-P., Vogel, G., Becker, C., Kothe, S., Rummel, U., and Ahrens, B.: 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, https://doi.org/10.1127/metz/2016/0537, 2016.

Sharma, A., Wuebbles, D. J., and Kotamarthi, R.: The Need for Urban-Resolving Climate Modeling Across Scales, AGU Advances, 2, e2020AV000271, https://doi.org/10.1029/2020AV000271, 2021.

Simón-Moral, A., Santiago, J. L., and Martilli, A.: Effects of Unstable Thermal Stratification on Vertical Fluxes of Heat and Momentum in Urban Areas, Boundary-Layer Meteorol, 163, 103–121, https://doi.org/10.1007/s10546-016-0211-4, 2017.

Slater, A. G., Schlosser, C. A., Desborough, C. E., Pitman, A. J., Henderson-Sellers, A., Robock, A., Vinnikov, K. Y., Entin, J., Mitchell, K., Chen, F., Boone, A., Etchevers, P., Habets, F., Noilhan, J., Braden, H., Cox, P. M., Rosnay, P. de, Dickinson, R. E., Yang, Z.-L., Dai, Y.-J., Zeng, Q., Duan, Q., Koren, V., Schaake, J., Gedney, N., Gusev, Y. M., Nasonova, O. N., Kim, J., Kowalczyk, E. A., Shmakin, A. B., Smirnova, T. G., Verseghy, D., Wetzel, P., and Xue, Y.: The Representation of Snow in Land Surface Schemes: Results from PILPS 2(d), Journal of Hydrometeorology, 2, 7–25, https://doi.org/10.1175/1525-7541(2001)002<0007:TROSIL>2.0.CO;2, 2001.

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

Stavropulos-Laffaille, X., Chancibault, K., Brun, J.-M., Lemonsu, A., Masson, V., Boone, A., and Andrieu, H.: 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, https://doi.org/10.5194/gmd-11-4175-2018, 2018.

Stavropulos-Laffaille, X., Chancibault, K., Andrieu, H., Lemonsu, A., Calmet, I., Keravec, P., and Masson, V.: Coupling detailed urban energy and water budgets with TEB-Hydro model: Towards an assessment tool for nature based solution performances, Urban Climate, 39, 100925, https://doi.org/10.1016/j.uclim.2021.100925, 2021.

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

Stewart, I. D. and Oke, T. R.: Local Climate Zones for Urban Temperature Studies, Bulletin of the American Meteorological Society, 93, 1879–1900, https://doi.org/10.1175/BAMS-D-11-00019.1, 2012.

Stretton, M. A., Morrison, W., Hogan, R. J., and Grimmond, S.: Evaluation of the SPARTACUS-Urban Radiation Model for Vertically Resolved Shortwave Radiation in Urban Areas, Boundary-Layer Meteorol, https://doi.org/10.1007/s10546-022-00706-9, 2022.

Takane, Y., Nakajima, K., and Kikegawa, Y.: Urban climate changes during the COVID-19 pandemic: integration of urban-building-energy model with social big data, npj Clim Atmos Sci, 5, 1–10, https://doi.org/10.1038/s41612-022-00268-0, 2022.

Taylor, K. E.: Summarizing multiple aspects of model performance in a single diagram, J. Geophys. Res., 106, 7183–7192, https://doi.org/10.1029/2000JD900719, 2001.

Tremback, C. J. and Kessler, R.: 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), 1985.

Varentsov, M., Samsonov, T., and Demuzere, M.: Impact of Urban Canopy Parameters on a Megacity's Modelled Thermal Environment, Atmosphere, 11, 1349, https://doi.org/10.3390/atmos11121349, 2020.

Velasco, E., Pressley, S., Grivicke, R., Allwine, E., Molina, L. T., and Lamb, B.: Energy balance in urban Mexico City: observation and parameterization during the MILAGRO/MCMA-2006 field campaign, Theor Appl Climatol, 103, 501–517, https://doi.org/10.1007/s00704-010-0314-7, 2011.

Velasco, E., Roth, M., Tan, S. H., Quak, M., Nabarro, S. D. A., and Norford, L.: The role of vegetation in the CO_2 flux from a tropical urban neighbourhood, Atmospheric Chemistry and Physics, 13, 10185–10202, https://doi.org/10.5194/acp-13-10185-2013, 2013.

Velasco, E., Perrusquia, R., Jiménez, E., Hernández, F., Camacho, P., Rodríguez, S., Retama, A., and Molina, L. T.: Sources and sinks of carbon dioxide in a neighborhood of Mexico City, Atmospheric Environment, 97, 226–238, https://doi.org/10.1016/j.atmosenv.2014.08.018, 2014.

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

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

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

Wang, Z.-H., Bou-Zeid, E., and Smith, J. A.: A Spatially-Analytical Scheme for Surface Temperatures and Conductive Heat Fluxes in Urban Canopy Models, Boundary-Layer Meteorol, 138, 171–193, https://doi.org/10.1007/s10546-010-9552-6, 2011b.

Wang, Z.-H., Bou-Zeid, E., and Smith, J. A.: 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, https://doi.org/10.1002/qj.2032, 2013.

Ward, H. C., Evans, J. G., and Grimmond, C. S. B.: 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, https://doi.org/10.5194/acp-13-4645-2013, 2013.

Ward, H. C., Kotthaus, S., Järvi, L., and Grimmond, C. S. B.: Surface Urban Energy and Water Balance Scheme (SUEWS): Development and evaluation at two UK sites, Urban Climate, 18, 1–32, https://doi.org/10.1016/j.uclim.2016.05.001, 2016.

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

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

Wouters, H., Demuzere, M., Blahak, U., Fortuniak, K., Maiheu, B., Camps, J., Tielemans, D., and van Lipzig, N. P. M.: The efficient urban canopy dependency parametrization (SURY) v1.0 for atmospheric modelling: description and application with the COSMO-CLM model for a Belgian summer, Geosci. Model Dev., 9, 3027–3054, https://doi.org/10.5194/gmd-9-3027-2016, 2016.

Yang, J., Wang, Z.-H., Chen, F., Miao, S., Tewari, M., Voogt, J. A., and Myint, S.: Enhancing Hydrologic Modelling in the Coupled Weather Research and Forecasting–Urban Modelling System, Boundary-Layer Meteorol, 155, 87–109, https://doi.org/10.1007/s10546-014-9991-6, 2015.

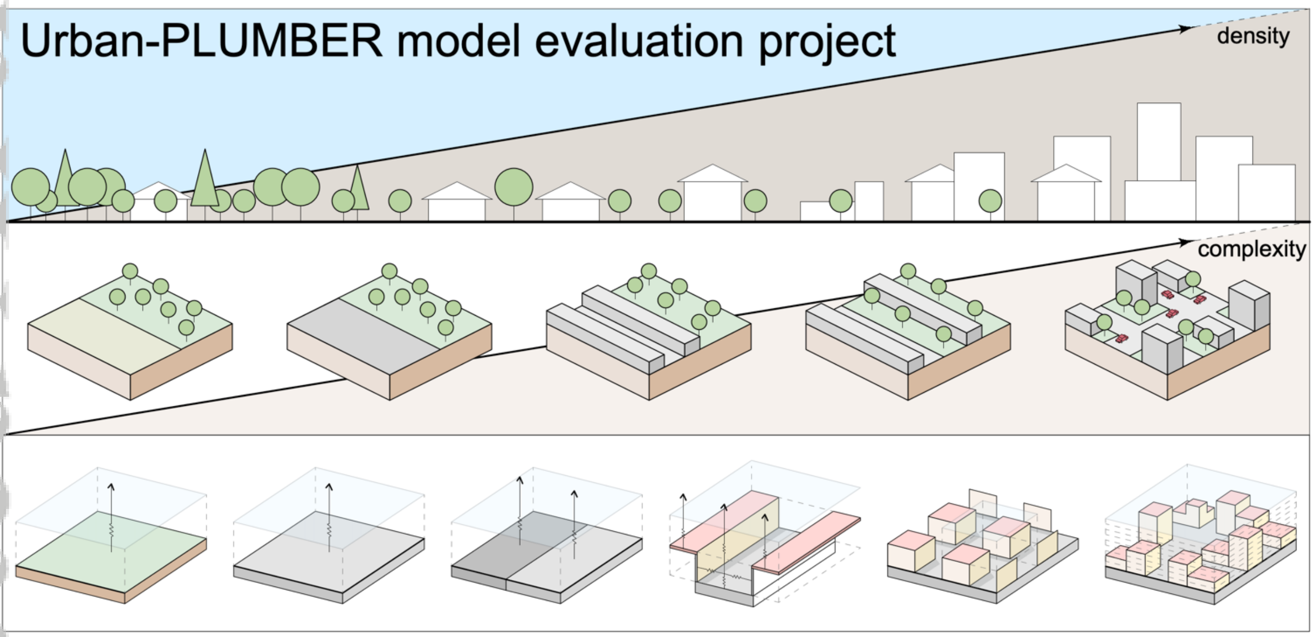
Yang, Z.-L., Dickinson, R. E., Henderson-Sellers, A., and Pitman, A. J.: Preliminary study of spin-up processes in land surface models with the first stage data of Project for Intercomparison of Land Surface Parameterization Schemes Phase 1(a), Journal of Geophysical Research: Atmospheres, 100, 16553–16578, https://doi.org/10.1029/95JD01076, 1995.

Zhao, L., Oleson, K., Bou-Zeid, E., Krayenhoff, E. S., Bray, A., Zhu, Q., Zheng, Z., Chen, C., and Oppenheimer, M.: Global multi-model projections of local urban climates, Nat. Clim. Chang., 11, 152–157, https://doi.org/10.1038/s41558-020-00958-8, 2021.

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

Mathew J. Lipson*, Sue Grimmond, Martin Best, Gab Abramowitz, Andrew Coutts, Nigel Tapper, Jong-Jin Baik, Meiring Beyers, Lewis Blunn, Souhail Boussetta, Elie Bou-Zeid, Martin G. De Kauwe, Cécile de Munck, Matthias Demuzere, Simone Fatichi, Krzysztof Fortuniak, Beom-Soon Han, Margaret A. Hendry, Yukihiro Kikegawa, Hiroaki Kondo, Doo-Il Lee, Sang-Hyun Lee, Aude Lemonsu, Tiago Machado, Gabriele Manoli, Alberto Martilli, Valéry Masson, Joe McNorton, Naika Meili, David Meyer, Kerry A. Nice, Keith W. Oleson, Seung-Bu Park, Michael Roth, Robert Schoetter, Andrés Simón-Moral, Gert-Jan Steeneveld, Ting Sun, Yuya Takane, Marcus Thatcher, Aristofanis Tsiringakis, Mikhail Varentsov, Chenghao Wang, Zhi-Hua Wang, Andy J. Pitman

*Corresponding author: Mathew J. Lipson (mathew.lipson@bom.gov.au)



We evaluate 30 land surface models' ability to simulate surface energy fluxes critical for meteorological and air quality simulations. Compared with the last major model intercomparison at the same site over a decade ago, we find broad improvement in the current cohort's predictions of shortwave radiation, sensible and latent heat fluxes, but little or no improvement in longwave radiation and momentum fluxes. We also conclude that human factors are likely to influence results in this (or any) model intercomparison.