Initial results from Phase 2 of the international urban energy balance model comparison

INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. **31**: 244–272 (2011)
Published online 7 October 2010 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/joc.2227

RMetS
Royal Meteorological Society

Initial results from Phase 2 of the international urban energy balance model comparison

C. S. B. Grimmond,[a]* M. Blackett,[a] M. J. Best,[b] J.-J. Baik,[c] S. E. Belcher,[d] J. Beringer,[e] S. I. Bohnenstengel,[d] I. Calmet,[f] F. Chen,[g] A. Coutts,[e] A. Dandou,[i] K. Fortuniak,[j] M. L. Gouvea,[a] R. Hamdi,[k] M. Hendry,[b] M. Kanda,[l] T. Kawai,[m] Y. Kawamoto,[n] H. Kondo,[o] E. S. Krayenhoff,[p] S.-H. Lee,[c] T. Loridan,[a] A. Martilli,[q] V. Masson,[r] S. Miao,[s] K. Oleson,[h] R. Ooka,[n] G. Pigeon,[r] A. Porson,[b,d] Y.-H. Ryu,[c] F. Salamanca,[q] G.J. Steeneveld,[t] M. Tombrou,[i] J. A. Voogt,[u] D. T. Young[a] and N. Zhang[v]

[a] *Department of Geography, King's College London, London WC2R 2LS, UK*
[b] *Met Office, FitzRoy Road, Exeter, EX1 3PB, UK*
[c] *School of Earth and Environmental Sciences, Seoul National University, Seoul 151-742, Republic of Korea*
[d] *Department of Meteorology, University of Reading, Reading, RG6 6BB, UK*
[e] *School of Geography and Environmental Science, Monash University, Melbourne, Vic, 3800, Australia*
[f] *Equipe Dynamique de l'Atmosphère Habitée Laboratoire de Mécanique des Fluides (UMR CNRS 6598) Ecole Centrale de Nantes, B.P. 92101, F-44321 NANTES Cedex 3, France*
[g] *Research Applications Laboratory, National Center for Atmospheric Research, Boulder, Colorado, 80307, USA*
[h] *Earth System Laboratory, National Center for Atmospheric Research, Boulder, Colorado, 80307, USA*
[i] *National and Kapodistrian University of Athens, Faculty of Physics, Department of Environmental Physics and Meteorology, Laboratory of Meteorology, Building Physics V, University Campus, 157 84 Athens, Greece*
[j] *Department of Meteorology and Climatology University of Lodz Narutowicza 88 Lodz Poland 90139*
[k] *Royal Meteorological Institute, Department II, section 3 Avenue Circulaire, 3, B-1180 Brussels, Belgium*
[l] *Department of International Development Engineering. Tokyo Institute of Technology,2-12-1-14-9, O-okayama, Megro-KU, Tokyo, Japan*
[m] *Research Center for Environmental Risk, National Institute for Environmental Studies, 16-2 Onogawa, Tsukuba-City, Ibaraki, 305-8506 Japan*
[n] *School of Engineering, The University of Tokyo, 7-3-1 Hongo, Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan*
[o] *Research Institute for Environmental Management Technology, National Institute of Advanced Industrial Sciecne and Technology,Tsukuba, Ibaraki, 305-8569, JAPAN*
[p] *Department of Geography, University of British Columbia, Vancouver, British Columbia, V6T 1Z2, Canada*
[q] *Department of Environment, CIEMAT, Madrid, 28040, Spain*
[r] *CNRM-GAME, Météo France/CNRS, Toulouse, 31057 Cedex 1, France*
[s] *Institute of Urban Meteorology, China Meteorological Administration, Beijing, 100089, China*
[t] *Meteorology and Air Quality Section, Wageningen University, P.O. Box 47, 6700 AA Wageningen, The Netherlands*
[u] *Department of Geography, University of Western Ontario, London ON N6A 5C2 Canada*
[v] *School of Atmospheric Sciences, Nanjing University 22 Hankou Road, Nanjing, 210093, China*

ABSTRACT: Urban land surface schemes have been developed to model the distinct features of the urban surface and the associated energy exchange processes. These models have been developed for a range of purposes and make different assumptions related to the inclusion and representation of the relevant processes. Here, the first results of Phase 2 from an international comparison project to evaluate 32 urban land surface schemes are presented. This is the first large-scale systematic evaluation of these models. In four stages, participants were given increasingly detailed information about an urban site for which urban fluxes were directly observed. At each stage, each group returned their models' calculated surface energy balance fluxes. Wide variations are evident in the performance of the models for individual fluxes. No individual model performs best for all fluxes. Providing additional information about the surface generally results in better performance. However, there is clear evidence that poor choice of parameter values can cause a large drop in performance for models that otherwise perform well. As many models do not perform well across all fluxes, there is need for caution in their application, and users should be aware of the implications for applications and decision making.

KEY WORDS urban climate; energy balance; surface atmosphere exchanges; land surface modelling; sustainable cities; radiation; turbulent heat fluxes; evaporation

Received 29 March 2010; Revised 18 August 2010; Accepted 21 August 2010

1. Introduction

Land surface models (LSMs) parameterize energy exchanges between the surface and the atmosphere for a wide range of different land surface types (e.g. deciduous trees, coniferous trees, grasses, bare soil, and urban). They provide the lower boundary conditions (fluxes) to meso- and global-scale atmospheric models and are forced with meteorology from the overlying atmospheric model. A wide variety of approaches are taken to model the influence of the underlying land surface type. To

* Correspondence to: C. S. B. Grimmond, Department of Geography, King's College London, London WC2R 2LS, UK.
E-mail: sue.grimmond@kcl.ac.uk

model the exchanges for an urban environment, LSMs range from a relatively simple representation of the urban environment as an impervious slab to models that take into account the 3D geometry of buildings with varying heights and material characteristics (Grimmond *et al*., 2009, 2010). During the process of simplification inherent to modelling, urban LSM (ULSM) developers have also chosen whether or not, for example, to include turbulent latent heat and/or anthropogenic heat fluxes. Increasing complexity, however, comes at the cost of both greater computational requirements and the number of parameters requiring specification. As even the most complex models do not include the complete specifications of all exchange processes, of interest is what level of improvement in performance, if any, is obtained with increased complexity.

Previously ULSMs have been evaluated individually against observational datasets of fluxes (e.g. Grimmond and Oke, 2002; Masson *et al*., 2002; Dupont and Mestayer, 2006; Hamdi and Schayes, 2007; Krayenhoff and Voogt, 2007; Kawai *et al*., 2009; Loridan *et al*., 2010a, 2010b; Porson *et al*., 2010). Although providing useful insights, these studies lack a structure that facilitates robust intercomparison. Here the principles of the project for intercomparison of land surface parameterization schemes (PILPS) (Henderson-Sellers *et al*., 1993, 2003; Irranejad *et al*., 2003) are followed. This paper, the second in an international model comparison study (PILPS-urban), evaluates ULSM in a common and consistent manner. In the first paper (Grimmond *et al*., 2010), results from an evaluation that used a short dataset (14 days) for a known site were presented (hereafter called Phase 1). By knowing the site location, a modeller should be able to assign more appropriate parameter values. Here the results from a comparison of 32 urban LSM (Table I), which represent a range of approaches (Figure 1), are analysed for a longer dataset (16 months) with the participants initially not knowing the location of the site beyond its designation as urban (hereafter called Phase 2). The Phase 1 and 2 sites have very different land cover, most notably the amount of vegetation (less/more, respectively), and land use characteristics (industrial/residential, respectively). All participants in the second phase had to have completed Phase 1 (Grimmond *et al*., 2010); one model from Phase 1 is not part of Phase 2. Phase 2 was structured into four stages corresponding to the controlled release of information about the site to enable a comparison of the importance of the parameters for each of the models. Although each group is informed how their own model performs, each one is not told about individual performance of other models.

The objectives of this paper are

1. To evaluate the ability of ULSM, in general, to model urban energy balance fluxes when provided with varying degrees of information about the urban environment.
2. To evaluate the performance of models with similar characteristics and complexity.
3. To reveal opportunities for future improvement of ULSM.

The first objective aims to highlight what might be expected in terms of ULSM performance when modelling urban energy balance fluxes for an area when only limited information is available about the site. With a steady release of surface characteristics it is possible to assess what surface information is most critical for optimal model performance. With these results it is also possible to address the second objective, the results of which will aid users in assessing what type of modelling approach is most appropriate for further development or for a particular application.

2. Methodology

To participate in this comparison a model had to simulate urban energy balance fluxes from the forcing data provided (Table II). The urban energy balance for these purposes is defined as:

$$Q^* + Q_F = Q_H + Q_E + \Delta Q_S \quad (1)$$

where Q^* is the net all wave radiation flux density which consists of the incoming shortwave ($K_\downarrow$) and longwave ($L_\downarrow$) radiation, which was provided as part of the forcing data, and the outgoing shortwave ($K_\uparrow$) and longwave ($L_\uparrow$) radiation which have to be modelled as:

$$Q^* = (K_\downarrow - K_\uparrow) + (L_\downarrow - L_\uparrow) \quad (2)$$

The anthropogenic heat flux (Q_F) may be modelled, prescribed, or ignored. All models have to simulate the turbulent sensible heat flux (Q_H), but the turbulent latent heat flux (Q_E) is neglected by some (Figure 1). All models calculate the net storage heat flux (ΔQ_S). The advective flux is not included in the energy balance at this scale, although it does not mean that advection does not exist. The micro-scale advection should be included within the sub-grid surface flux parameterizations. At the meso-scale, the inter-grid variations would be resolved by the overlying atmospheric model. Here, the ULSM are run independently of any large-scale model (i.e. offline). This is to ensure that the model performance evaluates the ULSM and not any compensation occurring within a larger scale model. It also ensures that the atmospheric conditions are fixed and independent of larger scale model performance. Similarly, this comparison neither evaluates the facet or micro-scale energy balance fluxes nor the vertical profiles within the urban canopy of the mean meteorological variables that some of the models are capable of calculating. Here we discuss only the results for the directly observed fluxes, so the storage heat flux and anthropogenic heat flux are not discussed. These will be discussed in future papers.

To conduct this comparison, the principles of the PILPS are employed. At the completion of each of the

Table I. The number of versions of each model used in the comparison and number of groups using it.

Code	Model name	References	Versions	Groups
BEP02	Building effect parameterization	Martilli *et al.* (2002)	1	1
BEP_BEM08	BEP coupled with building energy model	Martilli *et al.* (2002); Salamanca *et al.* (2009, 2010); Salamanca and Martilli (2010)	1	1
CLMU	Community land model – urban	Oleson *et al.* (2008a, 2008b)	1	1
IISUCM	Institute of industrial science urban canopy model	Kawamoto and Ooka (2006, 2009a, 2009b)	1	1
JULES	Joint UK land environment simulator	Essery *et al.* (2003); Best (2005); Best *et al.* (2006)	4	2
LUMPS	Local-scale urban meteorological parameterization scheme	Grimmond and Oke (2002); Offerle *et al.* (2003); Loridan *et al.* (2010b)	2	1
NKUA	University of Athens model	Dandou *et al.* (2005)	1	1
MORUSES	Met Office reading urban surface exchange scheme	Harman *et al.* (2004a, 2004b); Porson *et al.* (2010)	2	1
MUCM	Multi-layer urban canopy model	Kondo and Liu (1998); Kondo *et al.* (2005)	1	1
NJU-UCM-S	Nanjing University urban canopy model-single layer	Masson (2000); Kusaka *et al.* (2001)	1	1
NJUC-UM-M	Nanjing University urban canopy model-multiple layer	Kondo *et al.* (2005), Kanda (2005a, 2005b)	1	1
NSLUCM/ NSLUCMK/ NSLUCM-WRF	Noah land surface model/single-layer urban canopy model	Kusaka *et al.* (2001); Chen *et al.* (2004); Loridan *et al.* (2010a)	3	3
SM2U	Soil Model for submesoscales (urbanized)	Dupont and Mestayer (2006); Dupont *et al.* (2006)	1	1
SNUUCM	Seoul National University urban canopy model	Ryu *et al.* (2009)	1	1
SRUM2/ SRUM4	Single column reading urban model tile version	Harman and Belcher (2006)	4	1
SUEB	Slab urban energy balance model	Fortuniak (2003); Fortuniak *et al.* (2004, 2005)	1	1
SUMM	Simple urban energy balance model for mesoscale simulation	Kanda *et al.* (2005a, 2005b); Kawai *et al.* (2007, 2009)	1	1
TEB	Town energy balance	Masson (2000); Masson *et al.* (2002); Lemonsu *et al.* (2004); Pigeon *et al.* (2008)	1	1
TEB-ml	Town energy balance with multi-layer option	Hamdi and Masson (2008); Masson and Seity (2009)	1	1
TUF2D	Temperatures of urban facets 2D	Krayenhoff and Voogt (2007)	1	1
TUF3D	Temperatures of urban facets 3D	Krayenhoff and Voogt (2007)	1	1
VUCM	Vegetated urban canopy model	Lee and Park (2008)	1	1

Note these are assigned anonymous numerical identifiers in the analysis.

four stages, additional site information was provided (Table II). In Stage 1, only the forcing data were provided along with knowledge that observations were measured at 6.25 times the mean roughness height (z_H) for an urban area. In later stages, more site information consisting of basic surface cover fractions (Stage 2), urban morphology (Stage 3), and characteristics of urban materials (Stage 4) was provided. From this information, further parameters could be derived by participants as necessary (Grimmond *et al.*, 2010). After the completion of each run, participants sent back the calculated fluxes and the parameter values used for their model runs.

The site selected for Phase 2 was chosen based on having (1) a year or more of data to allow seasonality to be incorporated into the modelling; (2) little previous use by modelling groups to test models; (3) an almost complete quality controlled dataset available (i.e. not just for certain meteorological conditions only); and (4) co-operation with those that were involved in the data collection to participate in PILPS-urban.

The Phase 2 observation site was in suburban (Preston) Melbourne, Australia (Coutts *et al.*, 2007a, 2007b). This location was concealed from participants until the completion of Stage 4 before which equivalent latitude

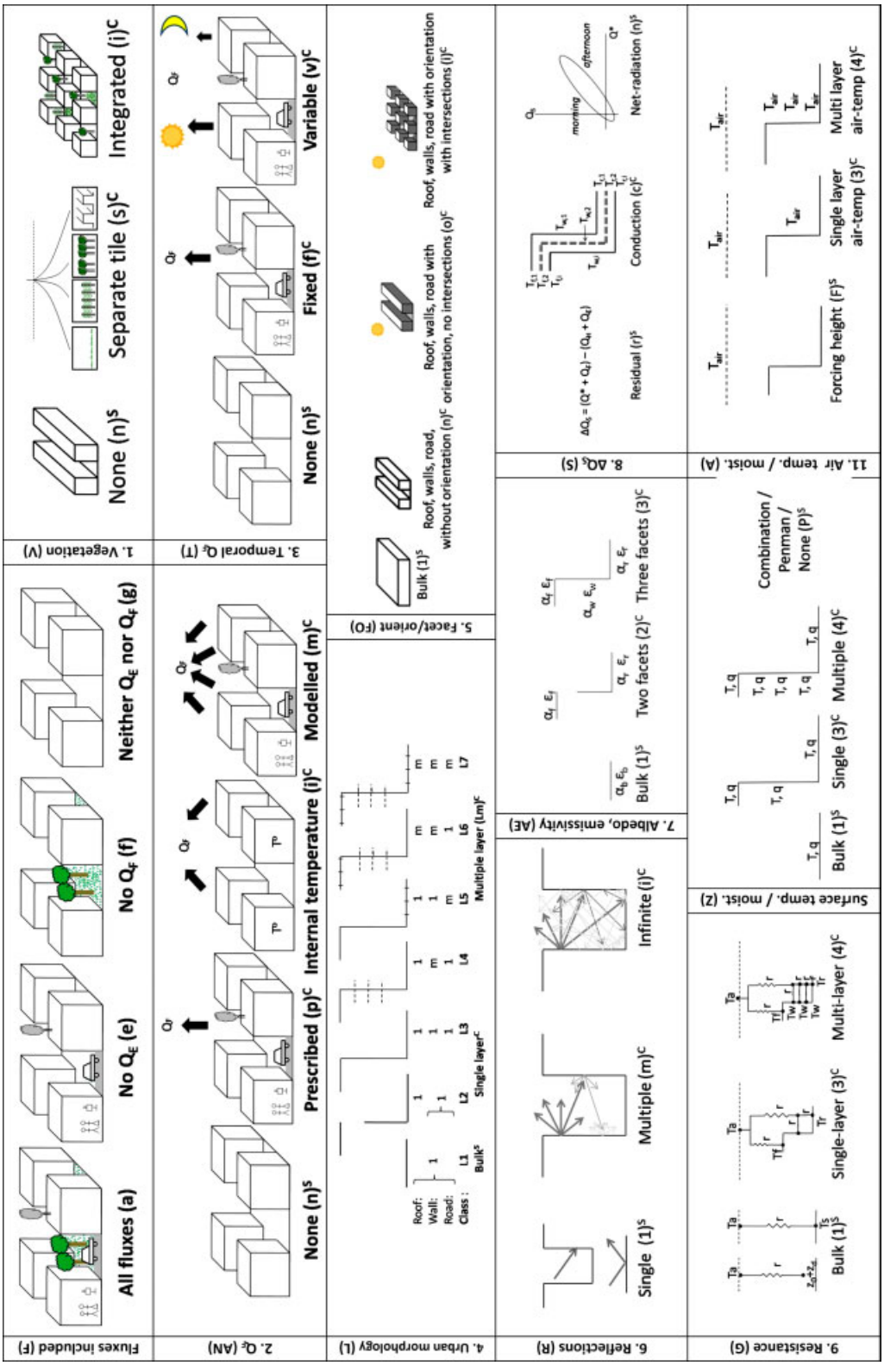

Figure 1. Model classifications with their individual characteristics. Each class characteristic is also classified as 'simpler' (S) or 'more complex' (C) (modified: Grimmond *et al.*, 2010). This figure is available in colour online at wileyonlinelibrary.com/journal/joc

Table II. Data provided at each stage.

	Category	Data provided		
Stage 1	Forcing data	K↓, L↓, air temperature, station pressure, specific humidity, wind components, rainfall		
	Site	Latitude*, Longitude*		
		Measurement height: 6.25 mean roughness height		
Stage 2	Plan area fraction	Pervious = 0.38		
		Impervious = 0.62		
Stage 3	Heights	Instrument height		40 m
		Roughness length for momentum		0.4 m
		Maximum height of roughness elements		12 m
		Mean building height		6.4 m
		Height to width ratio		0.42
		Mean wall to plan area ratio		0.4
	Plan area fraction	Surface cover	Fraction	Total
		Building	0.445	Impervious
		Concrete	0.045	0.62
		Road	0.130	
		Vegetation (excl. grass)	0.225	Pervious
		Grass	0.150	0.38
		Other (bare or pools)	0.005	
	Other	Urban climate zone = 5 Population density = 415.78 km^{-2}		

Stage 4: Material characteristics

Wall	d	C_p	c	λ
1	40.40	1008.5	1.25	0.61
2	54.00	1456.3	1.40	0.43
3	42.00	1010.0	0.0013[c]	0.024[d]
4	12.50	837.0	0.67[a]	0.16[a]
Roof				
1	11.6	865.2	2.07	6.53
2	50.00	965.3	0.0071	0.025
3	40.00	1880.0	1.50[a]	0.23[a]
4	12.50	837.0	0.67[a]	0.16[a]
Road				
1	28.75	912.79	1.14	1.17
2	158.3[a]	840	1.05[a]	0.30[a]
3	112.5[a]	840	1.29[a]	0.42[a]
4	650.45[a]	801	1.43[b]	3.72[b]

ρ[a] brick = 1500; softwood = 560.5; hardwood = 800; concrete = 1822; asphalt = 2100; glass = 2535.7; asbestos cement building board = 1920; asbestos cement tiles = 1900; terracotta = 1700; metal = 7900; fibreglass = 60; air = 1.29; gypsum/plaster board = 800; coarse crushed rock = 1250; fine crushed rock = 1540; sandy loam = 1780

Site albedo = 0.15 | Site emissivity = 0.973

Stage 4: details of layers components for each facet

	Wall						Roof						Road					
1	Material	%	C_p	c[a]	λ[a]	d	Material	%	C_p	c[a]	λ[a]	d	Material	%	C_p	c[a]	λ[a]	d
	Brick	27.94	840.0	1.26	0.71	110	Metal	30	1105	8.73	72.00	3.0	Asphalt	75	920	1.10	1.2	35
	Softwood	59.43	1975.5	1.11	0.14	20	Tile: Concrete	40	837	1.52	1.10	16.5	Concrete	25	837.0	1.5	0.87	10
	Concrete	5.37	837.0	1.52	0.87	100	Tile: Terracotta	20	837	1.42	0.99	16.5						
	Asbestos cement	5.37	1005.0	1.93	0.58	8	Tile: Asbestos cement	10	945	1.79	0.55	8.0						
	Concrete/wood	1.13	1406.2	1.31	0.79	60												
	Metal	0.76	1105.0	8.73	72.00	3												
2	Hardwood	80	1880	1.50	0.23	40	Air	85	1010	0.0013	0.024	50	Coarse crushed rock					
	Brick	20	840.0	1.26	0.71	110	Fibreglass	15	712	0.04	0.03	50						
3	Insulation (air)						Wood						Fine crushed rock					
4	Gypsum/plaster board						Gypsum/plaster board						Soil (sandy loam)					

The exact latitude and longitude (*) were not known only an equivalent for solar zenith angle is used. The material characteristics provided at Stage 4 consisted of information for four layers for each facet (roof, wall, and road) that included: layer composition/material, layer width (d, mm), specific heat capacity (C_p, J kg^{-1} K^{-1}) and volumetric heat capacity (c, MJ m^{-3} K^{-1}) which are related through density (ρ, kg m^{-3}) and thermal conductivity (λ, W m^{-1} K^{-1}) as well as the site observed mean albedo and emissivity.

[a] Clarke *et al.* (1991).

[b] Ochsner *et al.* (2001).

[c] Engineering Toolbox (2005a).

[d] Engineering Toolbox (2005b).

and longitude for solar zenith angle were released. The radiative fluxes were measured using Kipp & Zonen CM 7B and CG4 radiometers. Temperature and relative humidity were measured using a Campbell Scientific Inc. (CSI) HMP45C sensor. Both were sampled at 1 Hz and averaged to 30 min. To evaluate the modelled fluxes, the outgoing radiation components and its net balance were determined from Equation (2). The turbulent sensible and latent heat fluxes were measured using the eddy covariance technique. A CSI CSAT3 3D sonic anemometer was used with a CSI krypton hygrometer (KH20, August 2003 to February 2004) or a LI-COR LI7500 open path infrared gas analyser (February 2004 to November 2004). They were sampled at 10 Hz and block averaged using a CSI CR23X datalogger. The fluxes were calculated for 30-min intervals (Coutts *et al.*, 2007a, 2007b). Diurnal and seasonal Q_F fluxes were estimated for the site, following Sailor and Lu (2004); the estimates include sources of Q_F from vehicles, buildings (from the consumption of electricity and natural gas), and human metabolism (Coutts *et al.*, 2007b). The storage heat flux was calculated as the residual to Equation (1). This approach has the inherent problem that it accumulates all the measurement errors and missing terms (e.g. horizontal advection ΔQ_A) in this flux (Grimmond and Oke, 1999; Offerle *et al.*, 2005). However, Offerle *et al.* (2005) and Roberts *et al.* (2006) obtained close correspondence between fluxes from detailed facet temperature measurements and local-scale residual estimates of ΔQ_S. It is important to recognize that for all observations measurement errors occur. The observed fluxes and the forcing data are not without errors which are systematic and unsystematic. Typical errors are related to the instruments and their calibration, the meteorological conditions under which the observations are taken (e.g. changing turbulence conditions, shading), the processing of the data, the representativeness of the turbulent and radiant footprint, and siting of the instruments (Offerle *et al.*, 2003; Lee *et al.*, 2004; Hollinger and Richardson, 2005; Dragoni *et al.*, 2007; Foken 2008). Hollinger and Richardson (2005) have demonstrated that the sizes of uncertainty increase as a function of net all wave radiation therefore increasing with the size of the flux; for the growing season for Q_H and Q_E, uncertainty increases roughly as 0.1 Q^* and 0.08 Q^* when $Q^* > 0$ W m^{-2} and above ~10 W m^{-2} in an evergreen forest, respectively. They found no seasonality for Q_H errors but did for Q_E. Richardson *et al.* (2006) in an analysis of seven

sites, with a wider range of vegetation types, also found the error scales with the magnitude of the flux.

The forcing data consisted of 22 772 continuous 30-min intervals (474.5 days) from August 2003 to November 2004. Not all of the fluxes were available during all of these intervals, so here analysis is limited to the periods when all of the fluxes were measured. This gives 8865 intervals (38.9%) which were separated into two periods: the first 108 days and the last 365 days (8519 intervals when all fluxes were observed). The first period was to provide a spin-up, or initialization, period (the impact of this will be evaluated in a future paper). The post-initialization period allows for performance through an annual cycle to be evaluated.

Here, 32 different ULSMs are compared (Table I). The results are presented anonymously based on a randomly assigned unique model number. The models are grouped using a number of classifications based upon their characteristics (Figure 1) as described in Grimmond *et al.* (2010). To maintain anonymity, the number of models within each class had to be greater than three, thereby requiring some classes to be merged. Within each class, the approaches are categorized according to complexity (either simple or complex; Figure 1). Models are further categorized by their overall complexity depending on the number of 'complex' or 'simple' characteristics they possessed. The three groups are (1) 'complex' when all characteristics were complex (Cc), (2) 'medium' when the models possess one or two simple characteristics (Cm), and (c) 'simple' when they had three or more simple characteristics (Cs). Vegetation is not incorporated into this classification.

Comparison statistics reported here include root mean square error (RMSE), with both systematic ($RMSE_S$) and unsystematic ($RMSE_U$) components; the mean bias error (MBE); and the coefficient of determination (r^2). These are formally defined in Grimmond *et al.* (2010) from Willmott (1981) and Jacobson (1999). A larger systematic error typically indicates that the model has a problem in the model physics or parameter values, whereas a large unsystematic error is associated with the inability to cope with the variability in the observations which may be related to the 'randomness' of the conditions observed. Ideally, the systematic error would be the smaller of the two errors.

3. Results

3.1. Radiation comparison

To evaluate a model's ability to simulate radiative fluxes, the first aspect considered is whether there is closure in the radiation budget. Closure is assessed through comparison of the net all wave radiation (Q^*_{calc}) calculated from the two variables provided ($K_\downarrow$, $L_\downarrow$) and the two modelled variables ($K_\uparrow$, $L_\uparrow$) with the returned modelled Q^*_{mod}. No difference results in a coefficient of determination (r^2) of 1. At Stage 1, 15 of 32 models do not have a difference. In Stages 2/3/4, the number of models with $r^2 = 1$ is 13/16/13, respectively, but the total number of models that have $r^2 = 1$ at any stage is 18. Through four stages only ten models maintained no difference between Q^*_{calc} and Q^*_{mod}. If time periods with a difference of less than 1 W m^{-2} are considered (which includes one model with an r^2 of 0.999999), then Stages 1/2/3/4 have 16/14/16/13 models, respectively. These models are considered in the later analyses as being 'closed'. After this the r^2 values for Stage 1 range from 0.999991 to 0.0989 [*sic*]; with seven above 0.998, two more above 0.990, four more above 0.980, and two more greater than 0.870. The general groupings remain the same through the stages but the r^2 values do vary, except for the poorest models in Stages 1 and 2, which jump to greater than 0.998 at Stage 3.

Each modelling group which had a case of nonclosure was asked to determine the cause. The models without radiation balance closure problems are classified as P0 in the following analysis. Explanations for non-closure include (classified in analysis) not using the forcing data provided (P1), fluxes calculated independently (P1), timing issues (P3), day length (P3), spatial resolution (P3), and unknown (P4). In the first case, there are two different explanations: instead of using the individual 30-min interval forcing $K_\downarrow$ data, the daily peak observed $K_\downarrow$ was used and the other time periods for the day were obtained by assuming clear sky conditions, resulting in over-predicted $K_\downarrow$ and therefore Q^* (four cases, P1); and, the observed $L_\downarrow$ data were not used but modelled (one case, P1). In the second case, fluxes were calculated independently, the ULSMs calculate Q^* but for the purpose of this comparison, the radiative components have been calculated (three cases, P1) or there is an additional term in $L_\uparrow$ which is not incorporated into Q^* (one case, P4). In the third case, which relates to timing, the lack of closure is related to the 30-min forcing data being interpolated to a shorter time step for model calculations and then averaged back to the 30-min period for analysis (two cases, P3). This approach requires the forcing data to be interpolated which for $K_\downarrow$ may be questionable. For $L_\uparrow$, the approach depends on an emitted contribution from the surface temperature and a reflected part: $L_\uparrow(t) = (1-\varepsilon)L_\downarrow(t-\delta t) + \varepsilon\sigma {T_S}^4(t)$. The surface temperature T_S depends on the energy received and has inertia. Alternatively, it is because $K_\uparrow$ is only calculated if the sun is above the horizon for the whole time interval (one case, P3), thereby impacting the day length. The fifth case of spatial resolution (two cases, P3) is related to an underestimation of the total sky view factor (all model patches sum to less than 1.0) that arises in the process of rasterizing the surface within the model. The affected models then absorb slightly too much or too little diffuse solar or longwave radiation. The final case is where there are problems which the modelling groups have not been able to determine, leading to the imbalance (three cases, P4).

3.1.1. Outgoing shortwave radiation

The performance of each model, with respect to outgoing shortwave radiation ($K_\uparrow$), is shown in Figure 2 based on

RMSE; models that do not have closure are indicated. For this upwelling solar flux, only daytime fluxes are analysed. This gives 4266 × 30-min periods for comparison. The mean observed flux is 54.2 W m^{-2}. The Stage 1 $K_{\uparrow}$ mean RMSE for all ($N = 32$)/($N = 31$ models – as model 17 did not complete all stages)/not-closed/closed are 28/17/42/15 W m^{-2}, respectively, but the large difference is because of one model (17) which does not have closure. The mean RMSE for all 32 models by stage is generally larger than the median (Figure 2) because the mean is impacted by two poorly performing models, one of which did not complete Stage 4.

Considering all 32 models, as increasing information was provided (Stages 1–4) there was an improvement at each stage in mean but not in the median RMSE. The median RMSE improves from Stage 1 to 2 and again between Stage 3 and 4 (Figure 2). Of the 16/32 models with an improved RMSE from Stage 1 to Stage 2, 7/16 improved from Stage 2 to 3; and 2/7 of those improved from Stage 3 to 4. Thus, only two models had a reduction in RMSE at each stage. At Stage 2, improvement is associated with the fraction of vegetation to built areas becoming known (Table II). This fraction provides for the more realistic assignment of 'urban' and 'vegetated' albedos within the models. However, RMSE for five models became poorer. In Stages 2 and 3, a total of 14 models reduced (and 14 models increased) their RMSE and 13 in Stages 3 and 4 (and 4 increased). At Stage 3, more detailed information was provided about the surface fractions and heights. For the urban fraction it was now possible to distinguish the road and roof fractions correctly, in addition to knowing the wall heights. In the pervious fraction, grass could be distinguished from other vegetation. As expected, the largest overall improvement in $K_{\uparrow}$ based on the mean and median RMSE occurred at Stage 4 when the site observed albedo was provided (Figure 2).

The relative ordering of models in terms of performance remains relatively similar for all stages for $K_{\uparrow}$ with the same three models performing in the top three for all stages (Figure 2). Similarly, the poorest performing models, with slight reordering, remain the same for the four stages. But there are some notable changes for individual models between stages; e.g. model 22 does very well in Stages 1 and 2, then in Stage 3 the performance is much poorer but then returns to very good performance for $K_{\uparrow}$ in Stage 4. This demonstrates the importance not only of the model physics but also of the user's choice of parameter values, which can significantly influence the outcome. For Stages 1–3, there is a larger median systematic error ($RMSE_S$) than unsystematic error ($RMSE_U$), even when excluding model 17, but not for Stage 4 (Figure 2), suggesting that the additional surface information is important for improving the model performance. In Stage 4, once information about the albedo is available, 80% of the models have an $RMSE_U$ that is greater than the $RMSE_S$. The shading of the bars distinguishes the models complexity (C) among simple (s, yellow, light grey), medium (m, blue, medium grey), and complex (c, crimson, dark grey) (see Section 2 for definition). It can be seen that the three model types are distributed across the range of model performances, with all three occurring in the first and last five at Stage 1. By Stage 4, the Cc models are all in the middle group, (except a Cc model has dropped out). At Stage 4, the majority of the Cs models are doing well but the poorest performing model belongs to that group.

The effective albedo (α_{eff}) used in the models can be determined from $K_{\uparrow \text{mod}}/K_{\downarrow \text{obs}}$. Here this value is investigated at two times of the year (June 21 and December 21) at 13 : 00 h. These two times will have maximum and minimum amount of midday shadow. The range of values at Stage 1 is from 0.08 to 0.28 (except for two extreme outliers). The best performing model had an α_{eff} of 0.15, which was the same as the observed value provided at Stage 4, on both dates. The December 21 range of values were 4 (3) cases <0.1 (or >0.2); 3 (4) cases that were 0.10–0.125 (0.175–0.20); and 16 cases with an α_{eff} within 0.125–0.175, of which 11 have the lowest RMSE for $K_{\uparrow}$. For June 21, there was a similar distribution. The slightly higher α_{eff} (0.175–0.18) values are associated with the next best cohort in terms of RMSE performance.

The average cohort MBE is strongly influenced by the poorest performing models (Figure 2). The models have both positive and negative biases across the range which results in a net small negative bias (-4 W m^{-2} excluding model 17) for Stage 1. The median MBE has a large improvement from Stage 1 to 2 but after that remains almost constant at -1 W m^{-2}. At Stage 4, the Cm models which perform least well all have a negative bias, whereas the poor Cs models have both positive and negative MBE.

On a normalized Taylor (2001) plot, where the ideal model performance is indicated by the open circle at 1.0, 1.0, 0.0 (Figure 3), the correlation coefficient (polar) and normalized standard deviation (y-axis) and normalized RMSE (inner cirles) are shown. Except for one model the correlation coefficient is better than 0.8; for the majority of the models it is better than 0.9; and for many better than 0.99. One can track the impact of the additional information for the individual models; e.g. model 44 (medium complexity so blue with a symbol of a plus sign within a circle shown in Figure 3) in Stage 1 had a correlation of ~0.85 which improved in Stage 2 to ~0.91 and improved again in Stage 4 to ~0.95. Between Stages 2 and 3 there is a very minor change in correlation. In addition, one can see that there is an improvement in the normalized RMSE from greater than 0.5 to 0.4 to less than 0.4 (ideal is 0.0); and improvement of the normalized standard deviation from 0.62–0.73 to 0.74–0.88 (ideal is 1.0). For model 46 (same symbol but simple complexity) one can see that the model does not systematically improve.

Ensemble modelling, where the mean result from a number of different models is reported, is now used quite

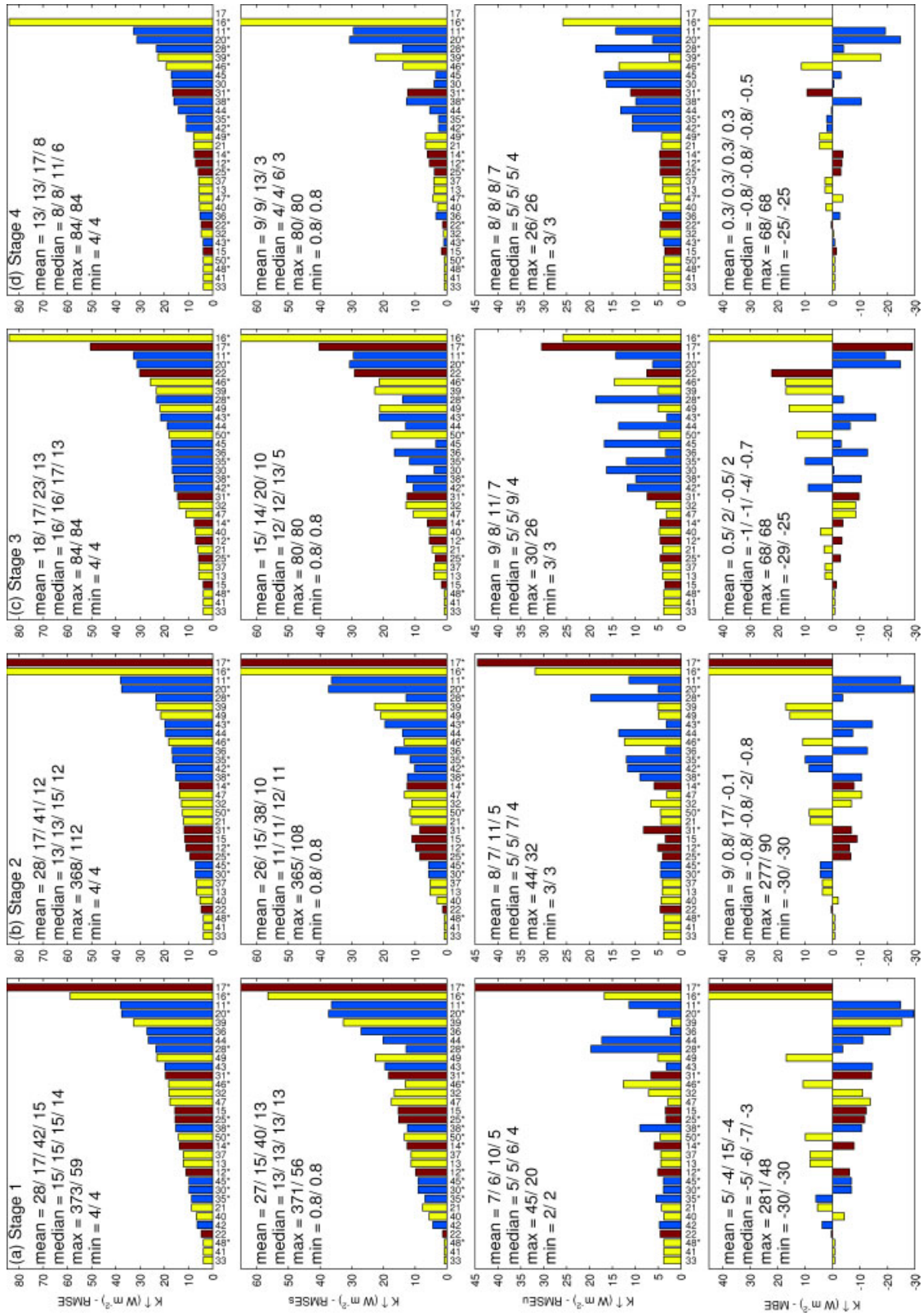

Figure 2. Model performance for each of the four stages (columns), for the last 12 months for outgoing shortwave radiation ($K_{\uparrow}$) (daytime only). The models are ranked based on RMSE for each stage with the systematic and unsystematic RMSE and MBE shown in the same order for each model. The overall statistics (mean, median, maximum, minimum) are given for the 32 models with and without model 17 ($N = 31$) for each stage in each figure. For the mean and median the statistics are also given for those models which do not have closure and do have closure of the radiation balance. The models which do not have radiative closure (see text) are indicated with a '*' (a–d). The shading of the bars distinguishes the models overall complexity (C) between simple (s, yellow, light grey), medium (m, blue, medium grey), and complex (c, crimson, dark grey) (see Section 2 for definition). The mean observed flux for this period was 54.2 W m^{-2}. This figure is available in colour online at wileyonlinelibrary.com/journal/joc

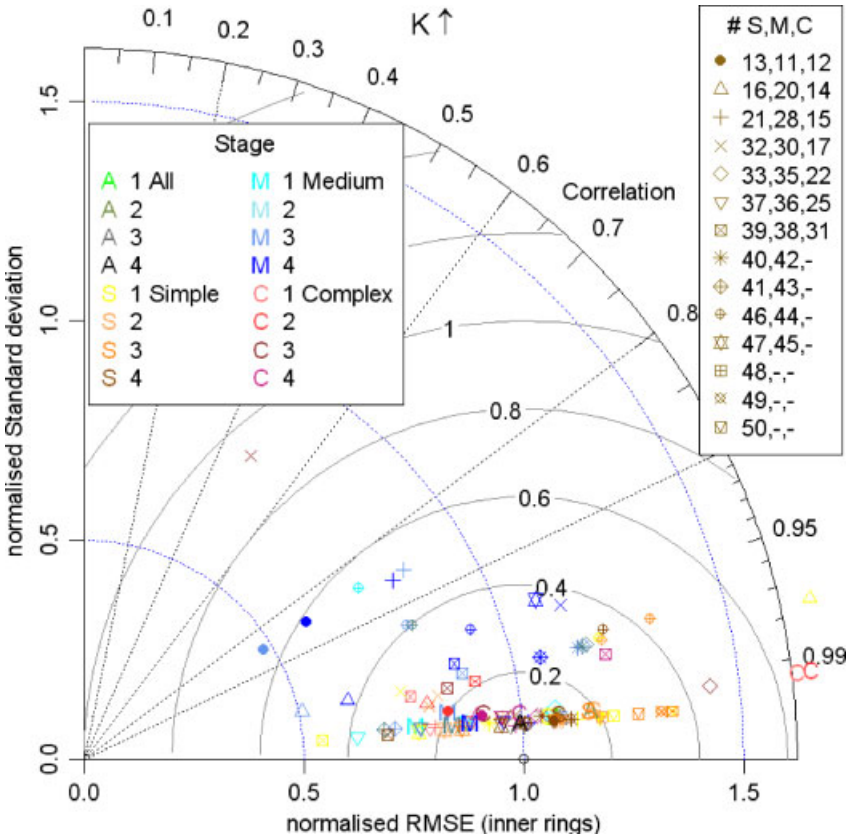

Figure 3. Normalized Taylor plot for the four stages, for the last 12 months for outgoing shortwave radiation ($K_{\uparrow}$) (daytime only). Taylor plots have the correlation coefficient on the polar axis, the normalized standard deviation on the radial *y*-axis and the normalized RMSE (*x*-axis) on the internal circular axes (Taylor, 2001). Performance for each model (symbol, colour indicates complexity and stage) and the ensemble results by complexity (letter) and stage (colour) are shown. Legend symbols are shown for simple Stage 4 colour. This figure is available in colour online at wileyonlinelibrary.com/journal/joc

extensively in the climate community (e.g. Gillett *et al.*, 2002). Here we consider the model performance of four different ensembles, three based on complexity [simple with 14 models, medium with 11, and complex with 7 (or 6 when model 17 drops out in Stage 4)] and the fourth is when all of the models are included [32 (Stages 1–3) or 31 (Stage 4)]. In Figure 3, these are shown for each stage. For the simple models, the correlation remains approximately constant, but there is an improvement in both the normalized standard deviation and normalized RMSE in the ensemble performance with stage. This is also the case for the medium and complex models. However, the ensemble performance of the complex models is clearly strongly influenced by the outlier model (17), which is beyond the plot boundaries, in Stages 1 and 2. At Stage 4 the ensemble performance is best when all (A) models are used but this is only slightly better than the ensemble mean performance of the complex models; the simple models' ensemble mean is slightly better than the medium complexity models.

The characteristics used to classify the models (Figure 1) include some that are directly related to radiative modelling. When the model results are grouped by these characteristics (Figure 4), we can determine if particular approaches result in better performance. In several classes, there is a clear separation in the mean performance associated with modelling $K_{\uparrow}$. However, in many cases the change in the mean is caused by one model's performance so the median is more robust as a measure of central location within the data. To maintain anonymity, each set of results plotted was required to have four or more results. This means that some classes are amalgamated. For each characteristic at each stage a box-plot of the RMSE gives the interquartile range (IQR), the individual models are plotted as dots, the median as a square, and the mean as a circle. Below each box the stage, the classification type, the characteristic with the class, then the number of models, the median, and the mean appear. For example, Figure 4(a) 1-Vn/11/14/17 indicates that for Stage 1 when the models are classified based on their approach to vegetation (V), there were 11 models that did not include it (n) with a median RMSE of 14 W m^{-2} and a mean of 17 W m^{-2}.

The first characteristic considered is whether the model integrates vegetation with the urban tile (Vi) rather than treating it separately (Vs) or not including it at all (Vn). For the Vi models there is a clear improvement in all four stages (Figure 4(a)). By Stage 3, the Vi models have a median RMSE of <4 W m^{-2} which is the smallest value. From Stage 2, when more models included vegetation (Vs models increase in number at

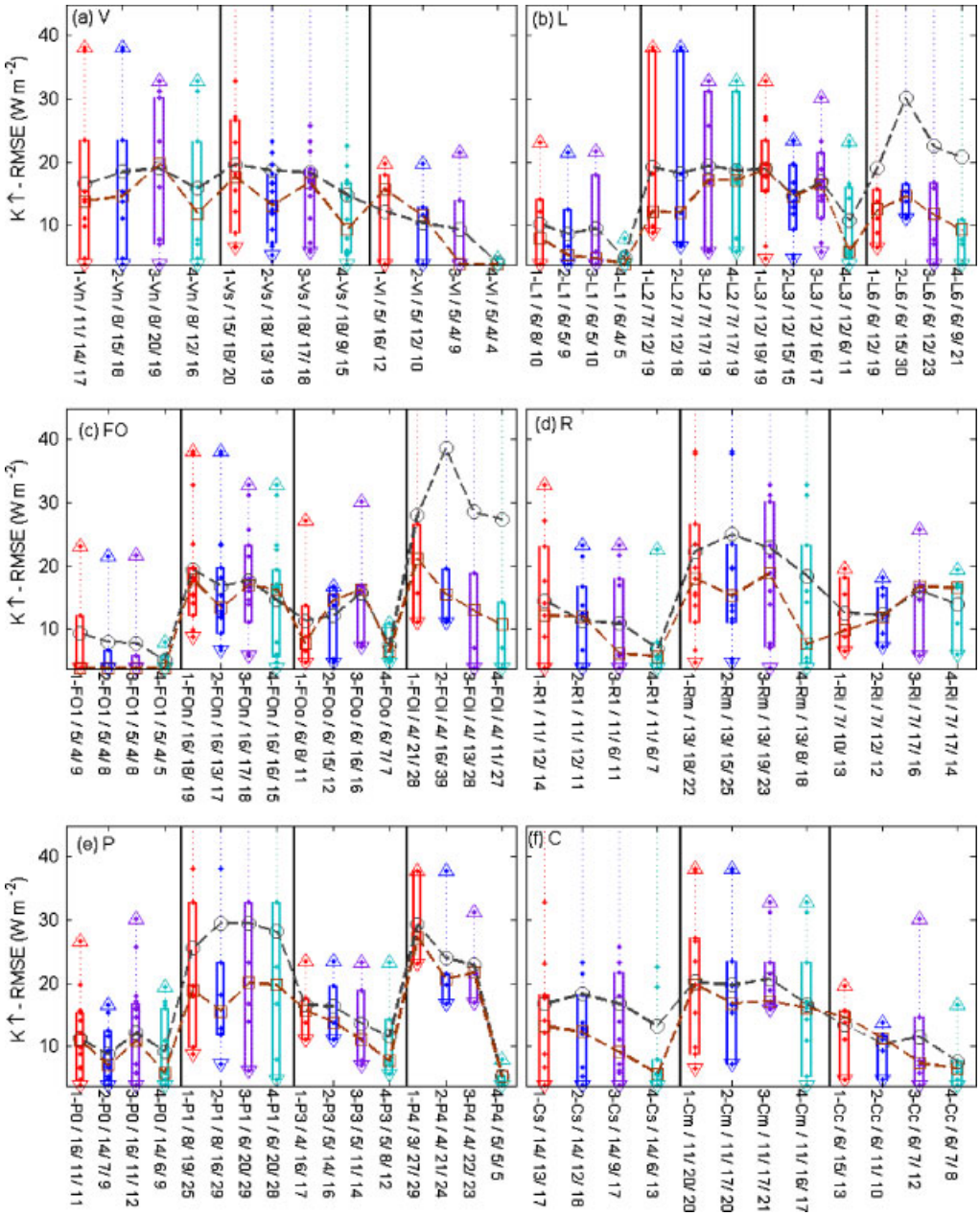

Figure 4. Model RMSE performance for the four stages, using the last 12 months for outgoing shortwave radiation ($K_{\uparrow}$) (daytime only) for the classes by approach taken (see Figure 1 for code interpretation or text): (a) vegetation (V), (b) urban morphology (L), (c) facets and orientation (FO), (d) reflections (R), (e) radiative closure performance (P), (f) complexity (C). Individual models are shown by the points, maximum and minimum by the triangles and the IQR by the box. Note the plots are cut-off at 0.40 of the maximum and the statistics are for $N = 31$ models (excludes 17). The circles are the mean of the cohort and the square is the median. The number of models, median, and mean are given for each. See text for further details. This figure is available in colour online at wileyonlinelibrary.com/journal/joc

the expense of Vn) the model cohorts retained the same ordering Vn, Vs, and Vi (decreasing median RMSE), but both Vn and Vs median performances deteriorate slightly in Stage 3. We can conclude that accounting for vegetation is important which is consistent with the conclusions from Phase 1 (Grimmond *et al.*, 2010).

Urban morphology (L) is specified using seven different approaches; from a slab surface (L1) to single-layer models (L2 – two components, L3 – three facets) and multi-layer (L4–7) models. The multi-layer models (L4–7) have different aspects of the surface that are treated in more detail (Figure 1) which leads to small numbers in each class. In this paper, these have been grouped together and labelled L6. This group has by far the largest mean RMSE because of one outlier (Figure 4(b)). The median performance for the simplest slab models (L1) improves at each stage and has the lowest median RMSE at Stage 4. For the other classes, there is not a consistent trend between stages; and for the L2 models the Stage 3 and 4 results have a higher median although reduced range, maximum and minimum than the earlier stages. The L3 models have second best median RMSE at Stage 4. Note for this characteristic that there

is no change in the model numbers per cohort between stages.

The approach to surface geometry with respect to whether the surface explicitly includes shaded surfaces or not (FO) has distinct differences between groups (Figure 4(c)). The simplest case, where the surface has a bulk geometry (FO1), has the lowest RMSE at all stages. It has a median RMSE of 4 W m^{-2} for all stages; however, the IQR decreases indicating more similar results. The most complex approach, which has both shading and intersections (FOi), has a systematic decrease in median RMSE at each stage, but at Stage 4 it is 11 W m^{-2}. This is greater than for models that take shading into account but have no intersection (i.e. have infinitely long canyons) (FOo) which have a median RMSE at Stage 4 of 7 W m^{-2}. The FOi models are clearly benefitting from the additional information provided, such as the wall height and built fraction provided at Stage 3. Both the FOo models and those that have an infinitely long canyon but do not account for shaded areas (FOn) have varying behaviour between stages; neither shows a continuous or significant improvement. The latter have the larger median RMSE at Stage 4 (16 W m^{-2}). The changing geometry influences the complexity of the modelling significantly with the simplest FO1 requiring considerably less computer resources than the more complete FOi which is theoretically much more realistic if within canyon information is required. Note, however, that the ability to model in-canyon information is not actually evaluated here.

Not only may the surface morphology description be different, but the approach taken to model reflections (R) also varies from those that include single (R1), multiple (Rm), or infinite reflections (Ri). The simplest (R1), unlike the other two approaches, has a systematic improvement in the median RMSE with stage (Figure 4(d)). By Stage 4, the median RMSE of 6 W m^{-2} is the smallest of the three approaches. The Rm approach, although it has a large scatter, shows a net improvement by Stage 4 (median RMSE = 8 W m^{-2}). The Ri group (median RMSE = 17 W m^{-2}) actually deteriorates through stages. So the simplest group consistently is the best performing and benefits from the additional information provided.

The albedo and emissivity (AE) classification distinguishes the amount of parameter information that is required by the models. The simplest case requires one bulk value (AE1) and so has a similar behaviour to FO1 and L1 (not shown). Significant improvement for these models at Stage 4 is a simple consequence of model formulation. Prior to Stage 4 albedo was assumed, but in Stage 4 for some models $K_{\uparrow}$ is just the product of two given values: site albedo and $K_{\downarrow}$. Models also can require two values (per parameter) typically associated with two facets (AE2) or three or more values (AE3). The median RMSE is lowest for the AE1 group and largest for AE2 (median RMSE at Stage 4 is 4 and 20 W m^{-2}, respectively). The vast majority of the models (22) require at least three values (AE3) for which the median RMSE by Stage 4 is 9 W m^{-2}; a net improvement from Stage 1. However this group, like the Rm, continue to have a wide range of values for the individual models.

The models that do not have a problem with net radiation balance closure (P0) have the smallest median RMSE at each Stage (Figure 4(e)). Their IQR does not have the smallest spread but the minimum values are lowest, and except for Stage 4, the 75 percentile is the lowest. The P3 (time and space resolution issues) and P4 models (unknown) have a systematic improvement with stage. At Stage 4, the median RMSE is 6/20/8/5 W m^{-2} for the P0/P1/P3/P4 models, respectively. The P1 models that have problems calculating a component of the radiative balance or did not use the forcing data for individual time intervals perform poorly throughout.

For all three model complexities, there are steady improvements in performance as additional information is provided (Figure 4(f)). The simplest and most complex (Cs, Cc) have a larger overall improvement than the Cm models with additional surface information. The Cs models have a slightly better median (6 rather than 7 W m^{-2}) but the mean is better for the Cc models (8 W m^{-2}).

Overall $K_{\uparrow}$ is modelled well and the provision of additional information about the surface does result in better performance. The models that perform best, for individual characteristics, are those that are the simplest as they can be assigned one parameter that is close to the observed value. The inclusion of vegetation is important to the performance. Based on overall complexity the simplest and the most complex models have similar results. The models that have net radiation closure perform better generally. The poorest performing cohort overall (P1) at Stage 4 does not have radiative closure and either did not make use of the individual time interval data and/or calculated the fluxes independently.

3.1.2. Outgoing longwave radiation

A combination of parameter information and flux calculations impact surface temperatures and hence the outgoing longwave radiation flux ($L_{\uparrow}$). Thus, the modelling of day- and night-time $L_{\uparrow}$ is more complex than modelling $K_{\uparrow}$ because of the relation between surface temperature, sensible heat, and storage heat fluxes, as well as $L_{\uparrow}$ itself. This means that, unlike the $K_{\uparrow}$ case, when additional information is provided more related parameters may be influenced.

For $L_{\uparrow}$, the median RMSE for the 32 models from Stages 1/2/3/4 are 16/13/14/17 W m^{-2}, respectively (Figure 5). Overall, 18 models improved from Stage 1 to 2, 11 from Stage 2 to 3, and 8 from Stage 3 to 4. Of the 32 models, only two improved across all the stages but eight improved in three consecutive stages. The largest improvement for an individual model was from Stage 2 to 3 with a greater than 20 W m^{-2} decrease in RMSE. The model performance from Stages 3 to 4, despite now having the most information about the site (Table II), suffered the largest loss of performance with 23 models

having an increase in RMSE. This relates to the trade-off that is made in parameter values. The largest individual performance deterioration also occurred between Stages 3 and 4 (increase of >35 W m^{-2} in the RMSE). There was one model that deteriorated across all four stages.

The models that close the radiation balance generally have better performance (e.g. smaller median RMSE) but that is not the case in Stage 1. At all stages, the models have a larger mean $RMSE_S$ than $RMSE_U$ but by Stage 3 and 4 the median $RMSE_U$ is slightly larger (Figure 5), suggesting that the model parameter information is appropriate for most of the models. In terms of the MBE more models have a positive bias rather than negative, but the two (one at Stage 4) models which perform least well have a large negative bias. The median MBE remains at about 8 W m^{-2} across all four stages.

The overall range of RMSE is smaller for $L_\uparrow$ than $K_\uparrow$ but the best performing model for $L_\uparrow$ has a larger RMSE than the best model for $K_\uparrow$. From comparison of the normalized Taylor plots (Figures 3 and 6), it is clear that the correlation is generally poorer for $L_\uparrow$. The mean $L_\uparrow$ flux is larger, but the diurnal range is smaller, than $K_\uparrow$. As with $K_\uparrow$, one (although different to $K_\uparrow$) model performs best across almost all stages (based on RMSE) and shows very little improvement with additional information being provided. This again is a simple model (Cs). The poorest performing model (excluding Model 17) does improve slightly with additional site information but still has a larger $RMSE_S$ than $RMSE_U$, suggesting that the model could be improved further. This differs from the next least well-performing model which has a larger $RMSE_U$ and a small positive MBE.

The three classes of complexity are scattered across the range of performance. However, again the best and poorest models are simple (Cs). In general, the simpler models are grouped in the middle or poorer end by Stage 4, whereas many of the Cm models are amongst the best. Unlike for $K_\uparrow$ the ensemble mean performance of the models does not improve with stage (Figure 6). At Stage 4 for all four ensembles all three measures have deteriorated. There is one model that is clearly performing better than the ensemble (but this is not the model with the lowest RMSE) pre-Stage 4. From the Taylor plot the best performing ensemble is the medium complexity but the four ensembles are clustered (and have moved together as a cluster between stages).

There is no model class that is better than the others. In most cases the model cohorts show poorest performance for all classes in Stage 4. For example (Figure 7), at Stage 4 the IQR is greater than in Stage 3 for all the approaches taken for vegetation (V); treatment of the urban morphology (L) has a drop in performance for each cohort in Stage 4, with the more complex models (L6) having the largest increase in median RMSE. There is very little change between stages in the other L classes. A similar result is obtained for the facet and orientation characteristics (FO) with no cohort improving across all four stages. One class (FOo) has a 6 W m^{-2} increase in median RMSE. For R and AE, similar results are obtained.

The models that have radiative closure (P0) have a median RMSE of 15 W m^{-2} at Stage 1 and 4. At Stage 4, the P0 cohort has the lowest median but this is not the case for all stages. For those without closure, the Stage 4 median is larger in all cases than Stage 1. For all P classes, Stage 2 was when the median RMSE was smallest.

The modelling of $L_\uparrow$ initially has the same size median RMSE as $K_\uparrow$ but not the general improvement with additional information (or progressive stage). This is seen consistently across all the classes of model types. In most cases, the Stage 4 results are poorer and have a larger IQR. At Stage 4, the best performing modelling approaches (lowest median RMSE) have the Vi, L3, FO1, Ri, AE1 and Cc characteristics. As was demonstrated previously (Grimmond *et al.*, 2010, Fig. 3), no single model has all these characteristics.

The models perform generally better at night than over the 24 h period (mean observed flux day = 410.14 W m^{-2} and night = 368.98 W m^{-2}). At night, the median RMSE for Stages 1/2/3/4 are 12/11/10/12 W m^{-2} and the median MBE are 8/7/2/−0.2 W m^{-2}. At Stage 4, the best performing (median RMSE W m^{-2}) models have Vn (13)/L2 (10)/FOn (11)/Rm (11)/AE2 (10) characteristics. Notably there is no difference between Cs/Cm/Cc models; they all have a median RMSE of 12 W m^{-2}. The daytime, as expected, is poorer with median RMSE for Stages 1/2/3/4 of 18/14/16/20 W m^{-2} and the median MBE are 9/7/9/12 W m^{-2}. At Stage 4, the best performing (median RMSE W m^{-2}) models have Vi(16)/L2 (17)/FOi (18)/Ri (15)/AE1 and AE3 (20)/Cc (15) characteristics. Thus, the characteristics that result in the lowest median RMSE change with time of day so there is not a clear choice, although the differences in the errors are small.

The models that do not have radiative closure occur across the complete spectrum of model performance for all time periods. The daytime median RMSE for P0 models improves from Stage 1 to 4 from 18 to 16 W m^{-2} but the Stage 2 result is the best for P0/P3/P4 models. For P1 models, the best performance is Stage 3 (15 W m^{-2}) but at Stage 4 the median RMSE is the poorest (26 W m^{-2}). At night the median RMSE for P0 models is 11 W m^{-2} at all stages (but deteriorating). The best performance is Stage 3/2/4 for P1/3/4 models.

Overall $L_\uparrow$ is not as well modelled as $K_\uparrow$. The daytime, when the mean flux is larger, has the larger median RMSE. The models generally improve when information about the pervious/impervious fraction is provided but generally did not improve when further details about heights and surface fractions were provided. Most models deteriorated when they were provided with details of the building materials typically back to Stage 1 performance but in many cases even poorer. Given both the wide range of materials that are in urban areas and the

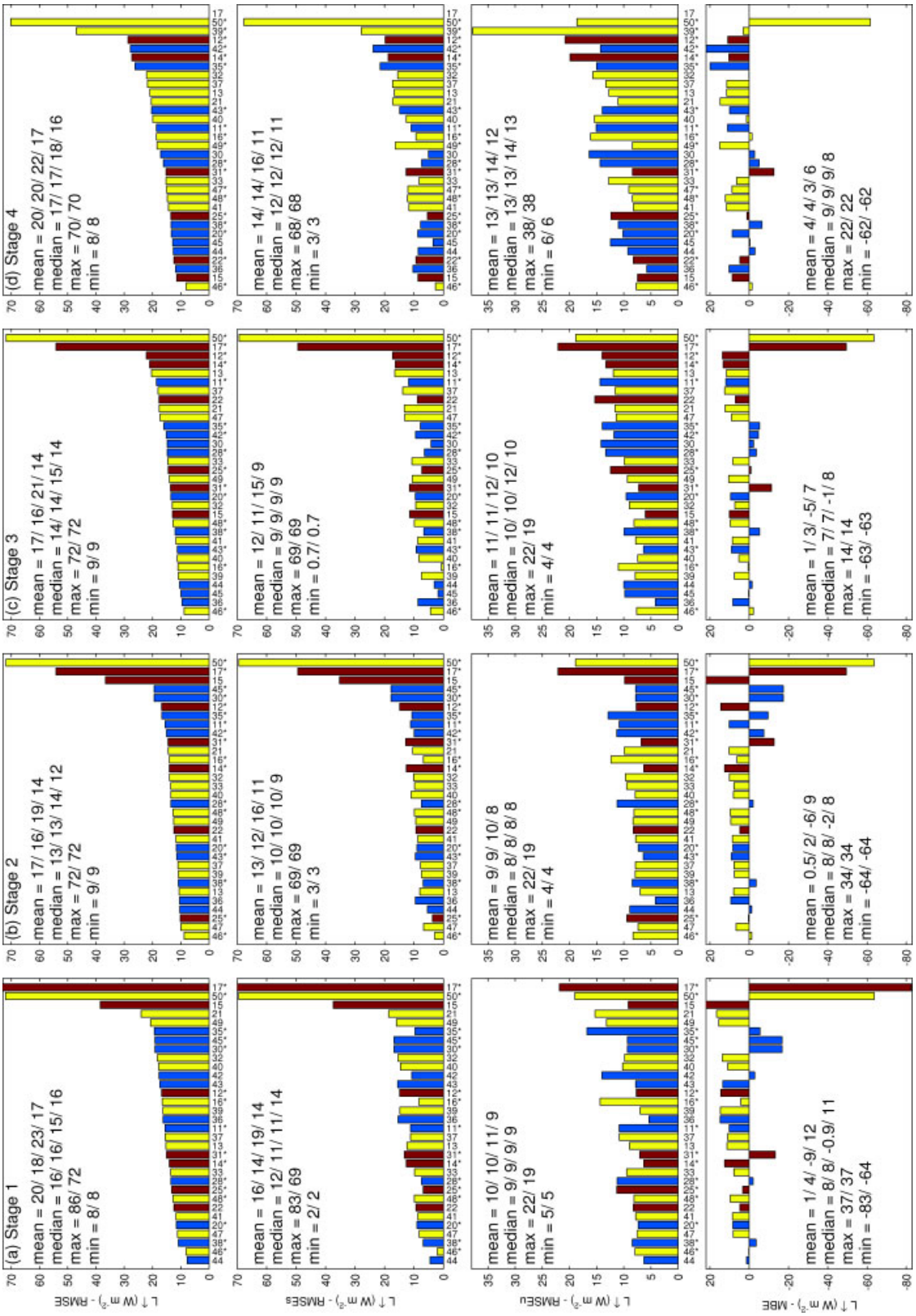

Figure 5. As for Figure 2 but for outgoing longwave radiation (L↑) for all hours. The mean observed flux for this period was 389.6 W m^{-2}. This figure is available in colour online at wileyonlinelibrary.com/journal/joc

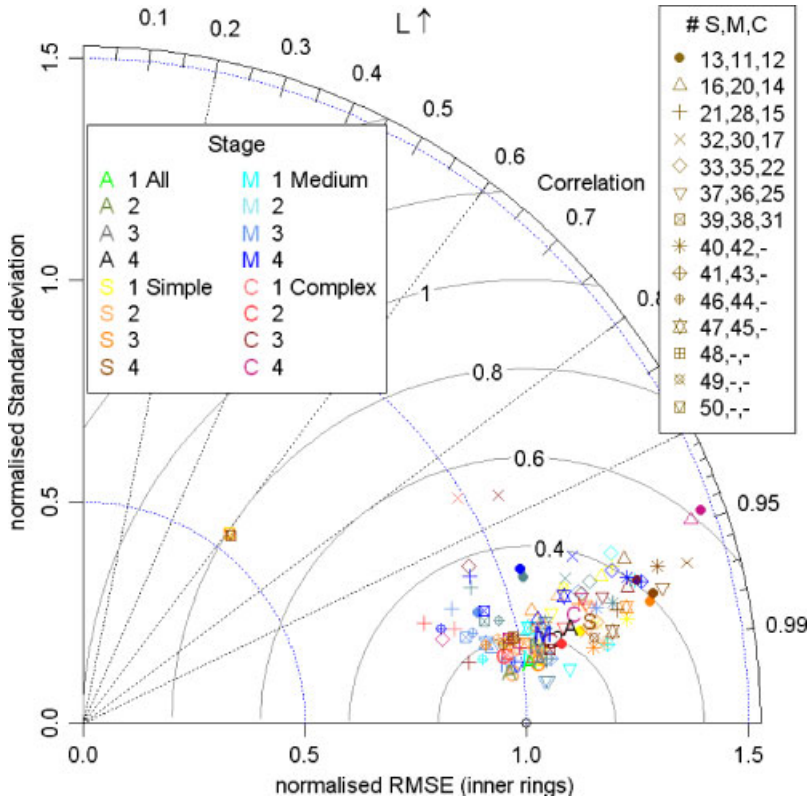

Figure 6. As for Figure 3 but for outgoing longwave radiation (L↑) for all hours. This figure is available in colour online at wileyonlinelibrary.com/journal/joc

associated wide range of values for individual material types, the difficulty of deciding what the appropriate values should be suggests that until there is a way to obtain realistic values for actual sites specification of materials may not be worth the effort required to obtain the information. Here we contacted a large number of people associated with the building and planning design plus materials suppliers (see 'Acknowledgements') to allow us to provide the data in Table II.

3.1.3. *Net all wave radiation*

Figure 8 shows the ranked performance of the models based on RMSE of net all wave radiation (Q^*), with the lack of radiative closure indicated. It can be seen from Figures 2, 5, and 8 that models which do not have closure are distributed from the best performing to the poorest performing for all three radiative fluxes evaluated, but are mainly the poorest performing for Q^*. For Stage 1 the mean RMSE for all models is 29 W m^{-2} for Q^* or 28 W m^{-2} when the model with poorest closure (r^2 of 0.0989) is removed because it did not complete all four stages. However, this model is not the poorest performing for Q^* but is for $K_{\uparrow}$ and $L_{\uparrow}$ at Stage 1 (Figures 2 and 5). Models that have radiative closure generally perform better over all stages for Q^* than those that do not; on average having a mean RMSE 20 W m^{-2} smaller. However, closure of the radiation balance is not a good measure of ability to calculate a particular flux. Comparing the performance of the components to the net all wave radiation shows a clear re-ranking between fluxes. Notably those that perform poorly for an individual component flux are not the poorest for Q^* (Figures 2, 5, and 8). This means that the application that the model is being used for is important; for example, when assessing a mitigation strategy's impact (such as changing the albedo of the materials on the change in radiative fluxes and temperatures) an ULSM may be modelling the most directly impacted flux well, but not able to model the other fluxes (or vice versa).

There were 14 models which showed a reduction in RMSE from Stage 1 to 2; of these five had a further improvement at Stage 3; and two of these improved again at Stage 4. However, in the opposite situation there are eight models whose RMSE increased from Stage 1 to 2; of which five had a further increase at Stage 3 and four had a further drop in performance at Stage 4.

The overall performance for Q^* does not vary much between stages though, with the mean RMSE being approximately 30 W m^{-2} at Stage 4, which is slightly larger than in the earlier stages. Also at Stage 4 models that do have closure of the radiation balance have a smaller mean and median RMSE (both 18 W m^{-2}, Figure 8). At Stage 4, however, these models have a slightly larger $RMSE_S$ than $RMSE_U$ suggesting that an improvement could still be made in the physics or parameter specification but this is not the case for both K↑ and L↑. The models generally have a negative MBE (Figure 8, Stage 4 median −6 W m^{-2}). The models with

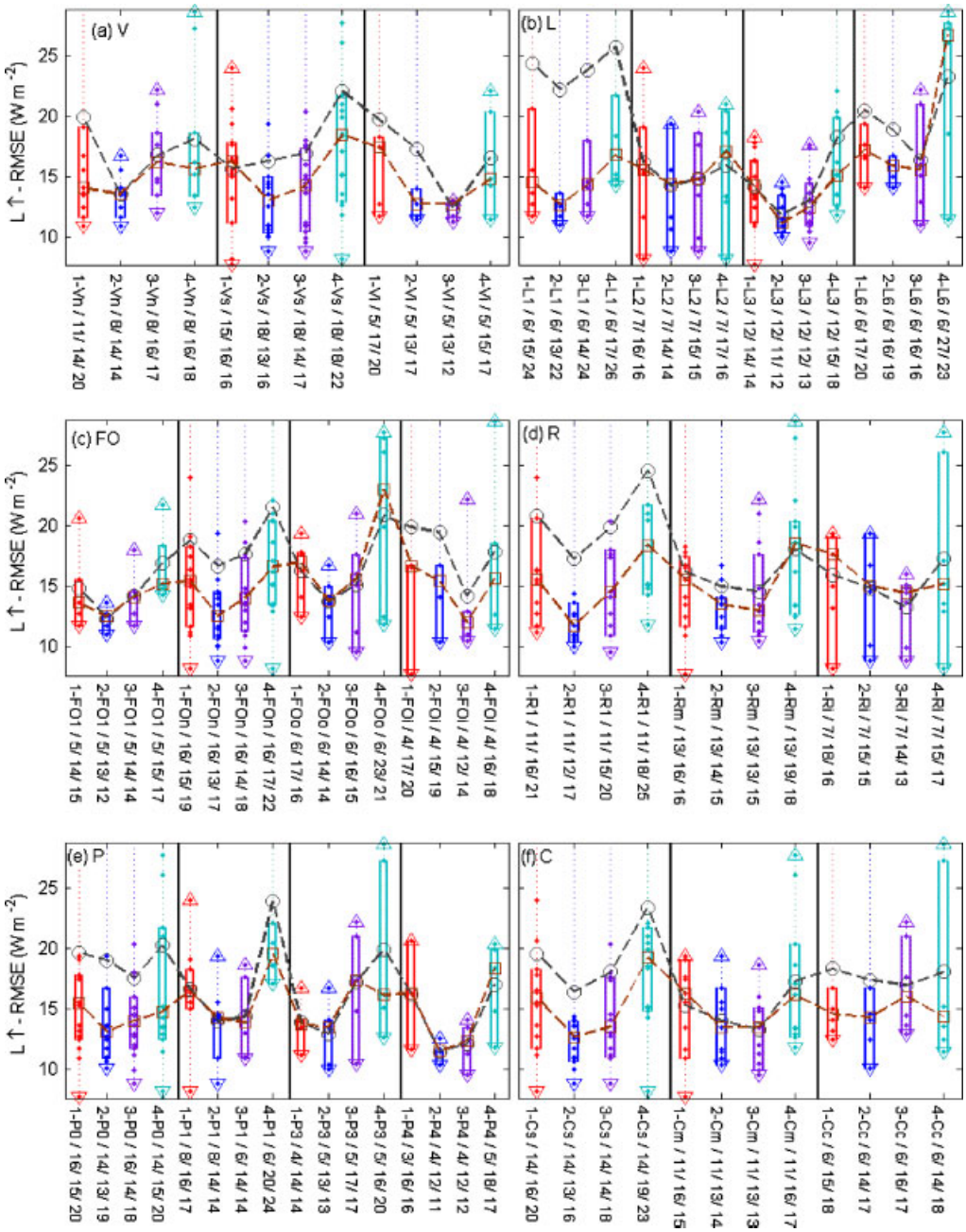

Figure 7. As for Figure 4 but for outgoing longwave radiation (L↑) for all hours. Note plots are cut-off at 0.40 of the maximum. This figure is available in colour online at wileyonlinelibrary.com/journal/joc

the largest absolute MBE are both positive and negative (Figure 8).

The best and poorest performing models at all stages are of medium complexity (Cm). At Stage 1 at both ends of the performance spectrum we have models from the three levels of complexity. By Stage 4 the more complex models have generally improved with three of the six (remember model 17 no longer appears) best performing models. Cm models are grouped more at the end with poor performance.

From the Taylor plot (Figure 9) it is clear, except for three models, all do an excellent job of modelling Q^*. There is a very tight cluster around (but not on) the ideal point. This performance is clearly better than for the separate radiative fluxes. Although this is good, this does suggest that there is some compensation occurring within the individual fluxes which may not be physically correct. As noted previously this result suggests that caution is needed when using the models to account for changing radiative characteristics. For the ensemble performance the medium complexity models are poorer than the other three. The best are the simple and complex models with slightly poorer performance from the 'all' ensemble.

The models that do not account for vegetation (Vn) show a steady decline in performance across all stages (Figure 10(a)). In contrast, there is no strong evidence

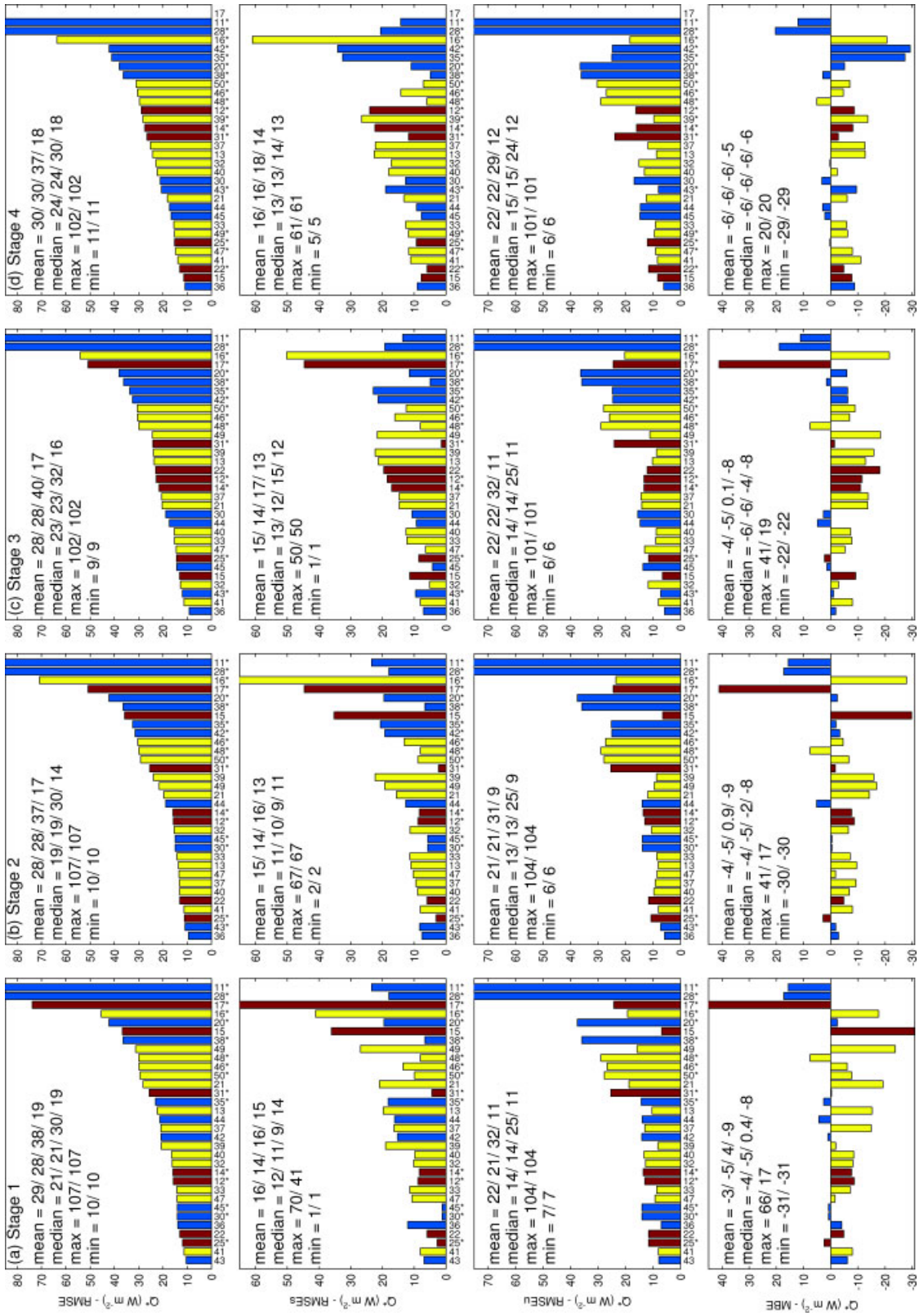

Figure 8. As for Figure 2 but for net all wave radiation (Q^*) for all hours. The mean observed flux for this period was 78.9 W m^{-2}. This figure is available in colour online at wileyonlinelibrary.com/journal/joc

for improvement by those that do include vegetation. The lowest median RMSE at Stage 4 (21 W m^{-2}) is for Vi models, but as for $L_{\uparrow}$, the performance deteriorates from 13 W m^{-2} at Stage 3. The best performing morphology class at Stage 4 is the simplest (L1) but the best performance across all stages and classes is Stage 2 L3, with a median RMSE of 14 W m^{-2}. This is the same result when the models are sorted by their approach to facets and orientation (FO) for the simplest models (FO1) at Stage 2, although FOo is only slightly larger at the same stage. This result is repeated again for classification based on treatment of R1 and for AE1.

The models with radiative closure (P0) have their lowest median RMSE at Stage 2 (15 W m^{-2}) and their largest at Stage 4 (25 W m^{-2}). The smallest median RMSE for P1 models is Stage 3 but these models have the largest IQR in Stages 3 and 4 (Figure 10(e)). As for $L_{\uparrow}$ at Stage 4, the complex (Cc) models perform slightly better than the less complex models even though they have deteriorated from better performance at earlier stages. The Cm models perform least well as a group with an increasing median RMSE with each stage.

The models perform generally better at night than for the 24 h period or for the daytime period (mean observed flux day = 216.83, night = −59.45 W m^{-2}). The night-time median RMSE for Stages 1–4 are 11/10/10/12 W m^{-2} and the median MBE are −7/−7/−2/1 W m^{-2}. At Stage 4, the best performing (median RMSE W m^{-2}) models have Vs (11)/L1&L2 (10)/FO1 (7)/R1 (7)/AE1 (7)/Cs (9) characteristics. The daytime performance for Stages 1–4 for the median RMSE was 27/24/28/29 W m^{-2} and for the median MBE was −5/−5/−8/−12 W m^{-2}. At Stage 4, the best performing (median RMSE W m^{-2}) models have Vi (28)/L1 (25)/FO1 (21)/R1 (25)/AE1 (21)/Cc (27) and Cs (28) characteristics. Compared to $L_{\uparrow}$ there is much greater variability between classes; e.g. the Cm models have daytime median RMSE of 50 W m^{-2} at Stage 4.

Models defined by simpler characteristics often perform best driven by the treatment of solar radiation. However, accounting for vegetation is important in improving the performance of the models. But when the overall complexity of the model is considered it is the more complex models that perform best overall and as a cohort make better use of the new site characteristics provided. The medium complexity models systematically drop in performance with increasing information provided, although there is consistently a Cm-type model performing best throughout.

3.2. Turbulent sensible heat flux

Model errors are larger for the turbulent sensible heat flux (Q_H) than for the radiative fluxes (compare Figures 2, 5, 8, and 11 and Figures 3, 6, 9, and 12). As for the radiative fluxes, the provision of information about the fraction of vegetation (Table II) results in an improvement with a reduction in median RMSE from 62 to 55 W m^{-2} (32 models). A similar sized reduction, down to 49 W m^{-2}, is evident at Stage 3, but at Stage 4 there is a small deterioration in performance (51 W m^{-2}). Throughout, the $RMSE_S$ is smaller than the $RMSE_U$, suggesting that overall RMSE is substantially driven by variability

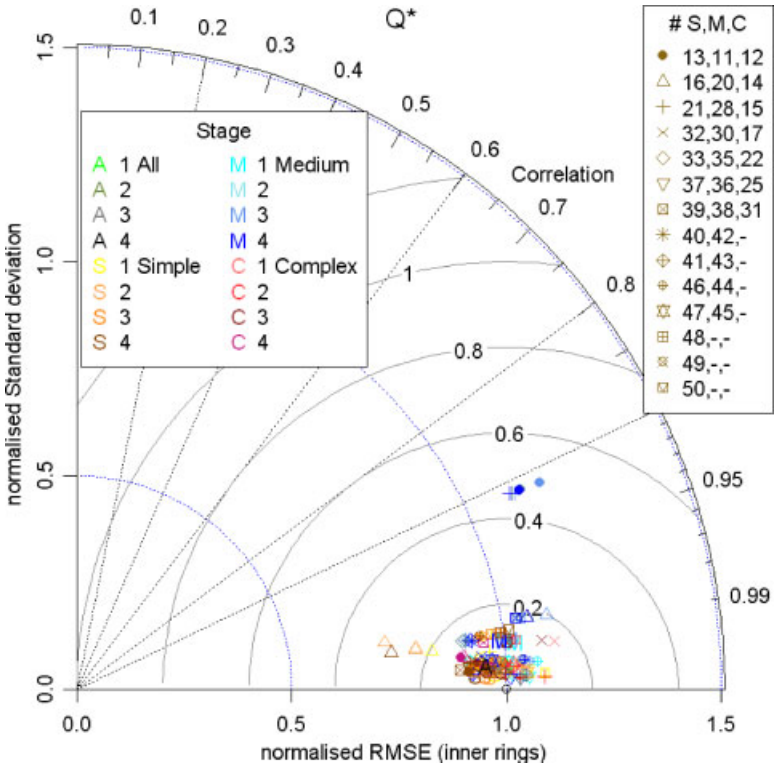

Figure 9. As for Figure 3 but for net all wave radiation (Q^*) for all hours. This figure is available in colour online at wileyonlinelibrary.com/journal/joc

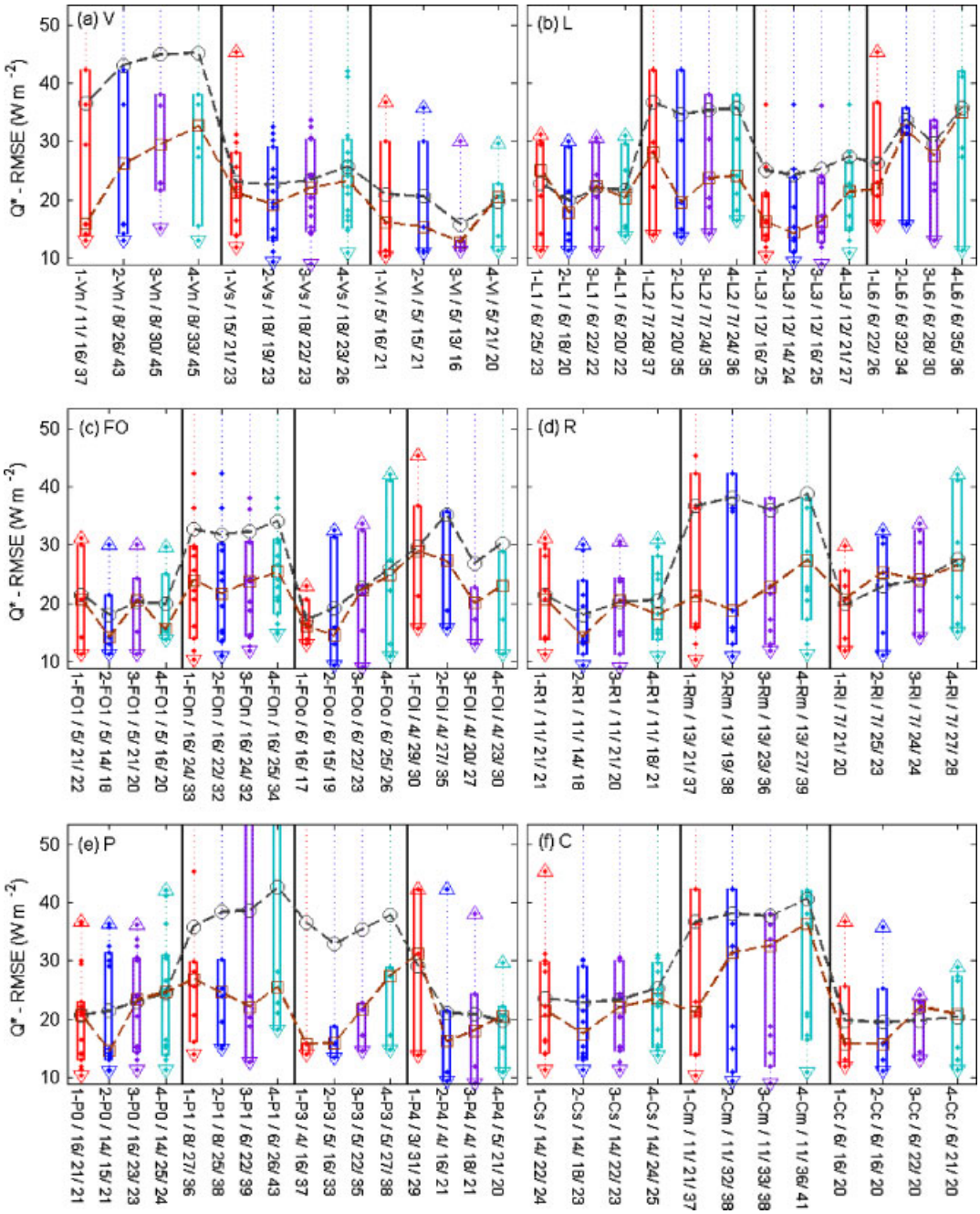

Figure 10. As for Figure 4 but for net all wave radiation (Q^*) for all hours. Note plots are cut-off at 0.50 of the maximum. This figure is available in colour online at wileyonlinelibrary.com/journal/joc

in the observed processes not included in the model physics and is less subject to improvement by better parameter specification. These may be at time scales the models do not capture. The median $RMSE_S$ drops (36/31/23/22 W m^{-2} – 31 models) at each stage as more information is provided about the site but unsystematic error remains around 42 W m^{-2} from Stage 2. The MBE is positive for most models and remains positive at all stages. The largest change in median MBE is at Stage 2, with a reduction from 20 to 6 W m^{-2}. In Stage 3, it rises slightly and then again at Stage 4. Overall there are five models with reduced RMSE at each stage (18 improved from Stage 1 to 2, 10 of which improved from Stage 2 to 3). There are also models whose performance deteriorates between stages; e.g. seven models from Stage 1 to 2 and of those two have a further increase in RMSE at Stage 3. From Stage 2 to 3, 11 models decline in performance (20 improved) followed by four which continue to increase their RMSE (10 improved) at the next stage. From Stage 3 to 4, 17 models improved (14 declined) in performance.

The model which performs best (or second best at Stage 4) is the model which did best for $K_{\uparrow}$, although it did not do best for Q^* or $L_{\uparrow}$. However, the daytime radiation should be reasonable because the shortwave dominates. The performance does not markedly improve through the stages for this model (i.e. there is not a

large reduction in the RMSE). At Stage 1, there are six models with RMSE which have a step drop in performance relative to the others (>10 W m^{-2}). None of these models have radiative closure. In the four stages the poorest model remains the same and has only a 7 W m^{-2} improvement as additional site data became available. Both the best and worst models in Stage 1 do not significantly improve by Stage 4, indicating that they are not benefiting from additional information. However, there is improvement within the middle range of models, most notably model 16 which performs best in Stage 4. The behaviour of the individual models with respect to systematic error shows some slightly surprising results. For example, model 50 which performs poorly overall has almost the smallest RMSE$_S$ overall. In fact, the small RMSE$_S$ are distributed throughout the range of the RMSE (Figure 11).

The normalized Taylor plots (Figure 12) show that no models or the ensembles have a correlation of 0.96 or greater. The ensemble mean values show generally better performance than the individual models. The ensemble of the simple models is the best with a clear change in performance between stages. When all models are included in the ensemble there is clear improvement from Stage 1 to 2 and 3 but reduced performance at Stage 4 (although it is slightly better than Stage 1). The modelling of Q_H is clearly poorer than the radiative fluxes and much more variable than for the radiative fluxes (e.g. compare Figures 9 and 12).

The models without radiation balance closure problems (P0) have a lower median RMSE than those that do not close (P1, P3, P4), except at Stage 4 (P4) (Figure 13(e)) when there is a rise in the median RMSE. For P1 (models which did not use the provided data) and P4 (unknown explanations), there is a reduction in RMSE across stages. Here we do not consider energy balance closure because the details of how Q_F enters the models are critical. Given the different assumption models made (Figure 1), it appears as an input, internal model assumption, and calculated output. At this stage we do not have all these values.

The impact of how vegetation is considered is seen clearly when comparing the Vn models to the Vs and Vi (Figure 13(a)). The Vn models have the widest range, largest IQR, and the poorest median performance. The Vi models perform the best but have a slight decrease in performance at Stage 4. The Vs cohort has the greatest improvement through the stages but also have a decrease at Stage 4. This suggests more complex and realistic treatments of vegetation may be important for modelling Q_H.

The simplest models with respect to morphology (L1) perform best relative to the others and improve across the stages (Figure 13(b)). The L2 models show the largest change between stages. The models which have a canyon but do not account for facet orientation (FO1) have the smallest median RMSE throughout and a steady reduction in the mean RMSE (Figure 13(c)). The treatment of surface temperature (Figure 1) for the built (B) fraction (Z_B) deteriorates with increasing complexity (not shown). The simplest (Z_B1) had an improvement at each stage with the median RMSE improving from 62 to 39 W m^{-2} across the four stages. In the other two approaches a steady improvement is not seen.

The treatment of AN varies from not including it or assuming it is negligible (ANn), to prescribing a value (ANp), to modelling it explicitly, or to using an internal temperature (ANc combined code of ANi, ANm, Figure 1). The simplest (ANn) has the lowest median RMSE and improves steadily across the four stages. Overall, the simplest models (Cs) have the smallest median RMSE at each stage, with improvements evident at each stage (Figure 13(f)). The median RMSE at Stage 4 for the three approaches with increasing level of complexity are 42/55/73 W m^{-2} (Cs/Cm/Cc). Thus, the simpler models often showed a net improvement with additional information, whereas that was not the case for the more complex models. This may be because there was not enough additional detailed information provided for the more complex models so it was more difficult for the users to decide how to use this information appropriately. In addition, such models typically have many more parameter values that could be altered in response to the new information provided.

The daytime results at Stage 1 have a larger median RMSE than the 24 h or night-time (79/62/28 W m^{-2}) which continues to Stage 4 (68/51/21 W m^{-2}). Obviously, the variability and the magnitude of Q_H is much greater during the daytime than for night-time hours (mean observed flux: day = 88.72, night = −13.16 W m^{-2}). The median daytime MBE is positive during the day (40/25 W m^{-2} Stage 1/4) and negative at night by Stage 4 (10/−8 W m^{-2} Stage 1/4). At night, there is one poor model (17) for the three stages, but there is another model that performs very poorly at Stage 3 but in Stage 4 returns to much better performance. These individual model RMSE results are >115 W m^{-2} compared to under <50 W m^{-2} for the remainder of the models. The poorly performing models during the daytime are different and the same two models perform poorly throughout (the difference to the next models is of the order of 50 W m^{-2}). Thus, the models that are performing least well on the all hour basis are caused by different abilities related to day- and night-time processes.

Overall, the simple complexity (Cs) models perform best but it is important to include vegetation. With additional information the models improve but the simplest models have a systematic improvement at each stage, whereas for the more complex models this is not the case. In this case, where Q_F is not very large, the models that do not account for Q_F do better. The slab or bulk models also show a consistent improvement at each stage.

3.3. Turbulent latent heat flux

The modelling of latent heat flux (Q_E) needs to deal with the loss of water from a wet surface, e.g. after rainfall from roofs, roads, and vegetation; and the transpiration

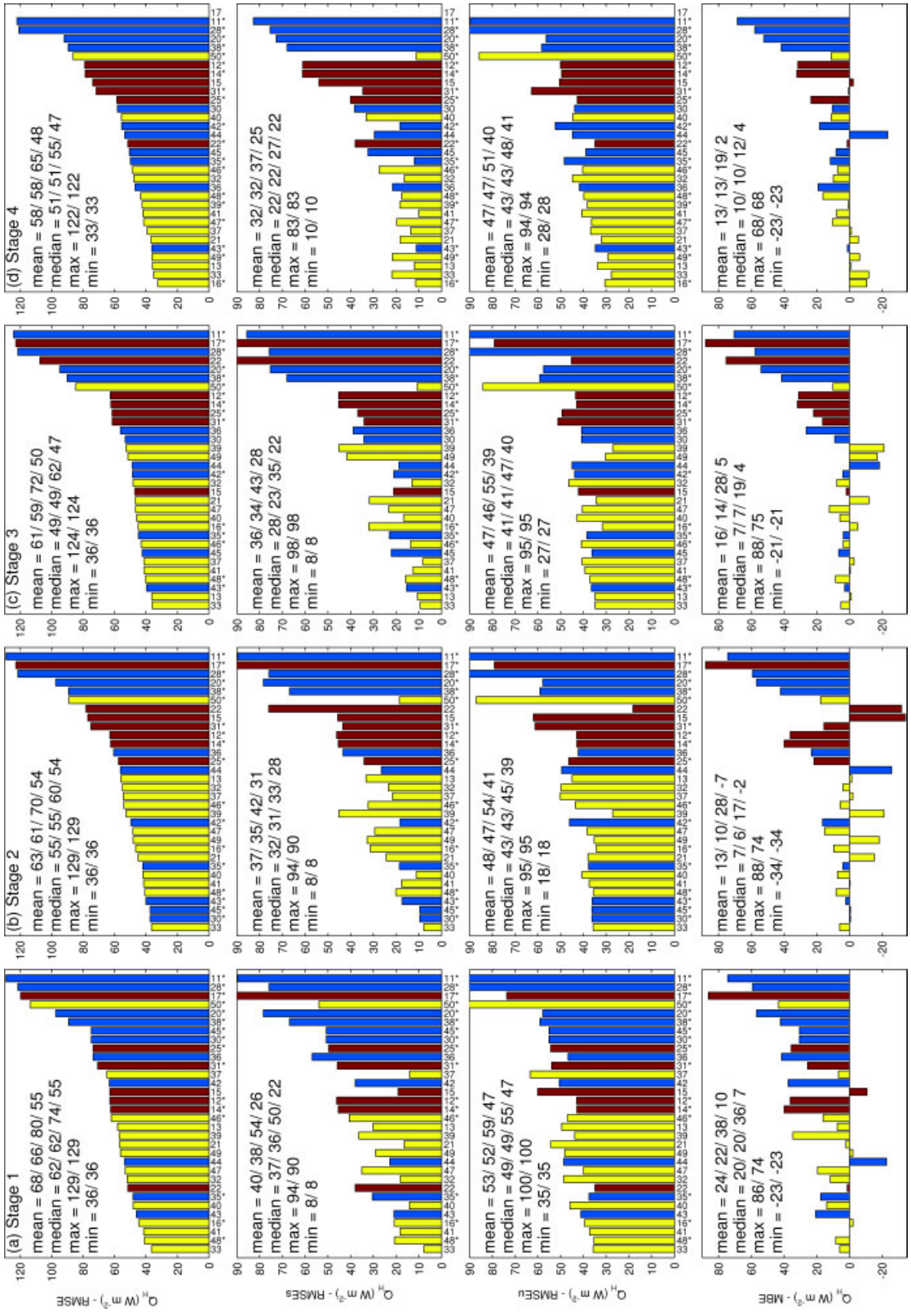

Figure 11. As for Figure 2 but for turbulent sensible heat flux (Q_H) for all hours. The mean observed flux for this period was 37.9 W m^{-2}. This figure is available in colour online at wileyonlinelibrary.com/journal/joc

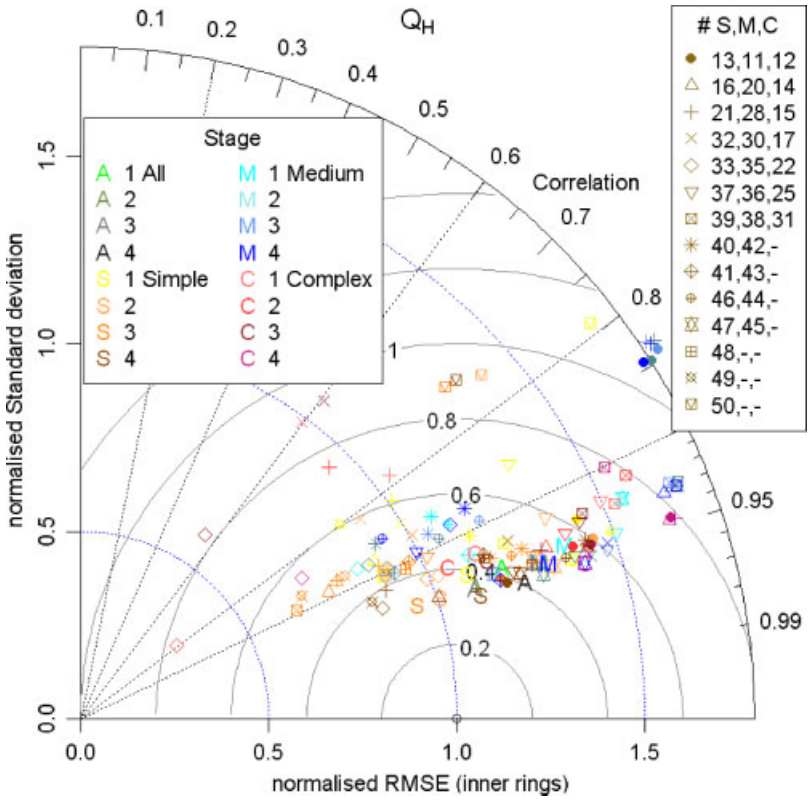

Figure 12. As for Figure 3 but for turbulent sensible heat flux (Q_H) for all hours. This figure is available in colour online at wileyonlinelibrary.com/journal/joc

of vegetation which continues between rainfall events. The median RMSE for the modelled latent heat flux (Figure 14) dropped by the largest amount at Stage 2 when information about the vegetation was provided (54/42/42/43 W m^{-2}, for 31 models). There was no general improvement from knowing more details about the plan area fractions of vegetation (e.g. grass *vs* nongrass, Stage 3). Across the four stages there are six (seven Stage 1) models that have a large $RMSE_S$ (58 W m^{-2}) and a 0 W m^{-2} $RMSE_U$; these are ignoring latent heat flux completely. There are a couple of models that address some aspect of this flux but have even poorer performance than those that neglect it. However, all but one of these models improves so by Stage 4 there is only one model that is in this category. It should be noted that this model does not close the radiation or the energy balance.

From Stage 1 to 2, 17 models have a reduced RMSE; 11 of which improve at Stage 3; and of these, four improve at Stage 4. In the reverse direction, of the eight models which have an increase in RMSE at Stage 2; three have a further increase at Stage 3 and one deteriorates again at Stage 4. Similarly, there is one model that has the largest increase in $RMSE_S$ at Stage 2 and retains this across the stages.

Overall, the systematic errors are generally larger than the unsystematic errors. As noted above, this is largely due to the models not attempting to model latent heat flux (Figure 14). By Stage 4, the median $RMSE_S$ has dropped by nearly 20 W m^{-2}, whereas the $RMSE_U$ remains about the same so there is a definite benefit from the new information provided (either directly as parameters or recognizing the need to consider particular processes more fully). Overall there is a negative MBE, with a median of -18 W m^{-2} at Stage 1. The best performing models based on MBE at Stage 1 have a small positive MBE but the majority have a negative MBE. By Stage 2, the MBE halved to -9 W m^{-2}. This obviously remains large because of those models that have not modelled Q_E but does suggest that those that do include it are generally underestimating the flux. This could be because they do not account for additional urban sources of water through irrigation, which can influence evaporation rates and soil moisture (Grimmond and Oke, 1991). This information was not provided at any of these stages to the model participants.

Initially, except for one Cc model, all the best performing models are simple models and the Cm are all grouped at the poorer performing end (Figure 14). However, at Stage 2, when vegetation fraction became known, Cm models start to improve. By Stage 4, we have all model types represented at the poor end, but the five models with the lowest RMSE are Cs.

The correlation coefficient for all models at all stages is less than 0.8 (Figure 15). This result along with the other normalized statistics on the Taylor plot, demonstrates that Q_E is the least well-modelled flux (compare Figures 3, 6, 9, 12, and 15). There is even wider scatter amongst the models than for Q_H. The ensemble performances generally have the better correlations but the normalized

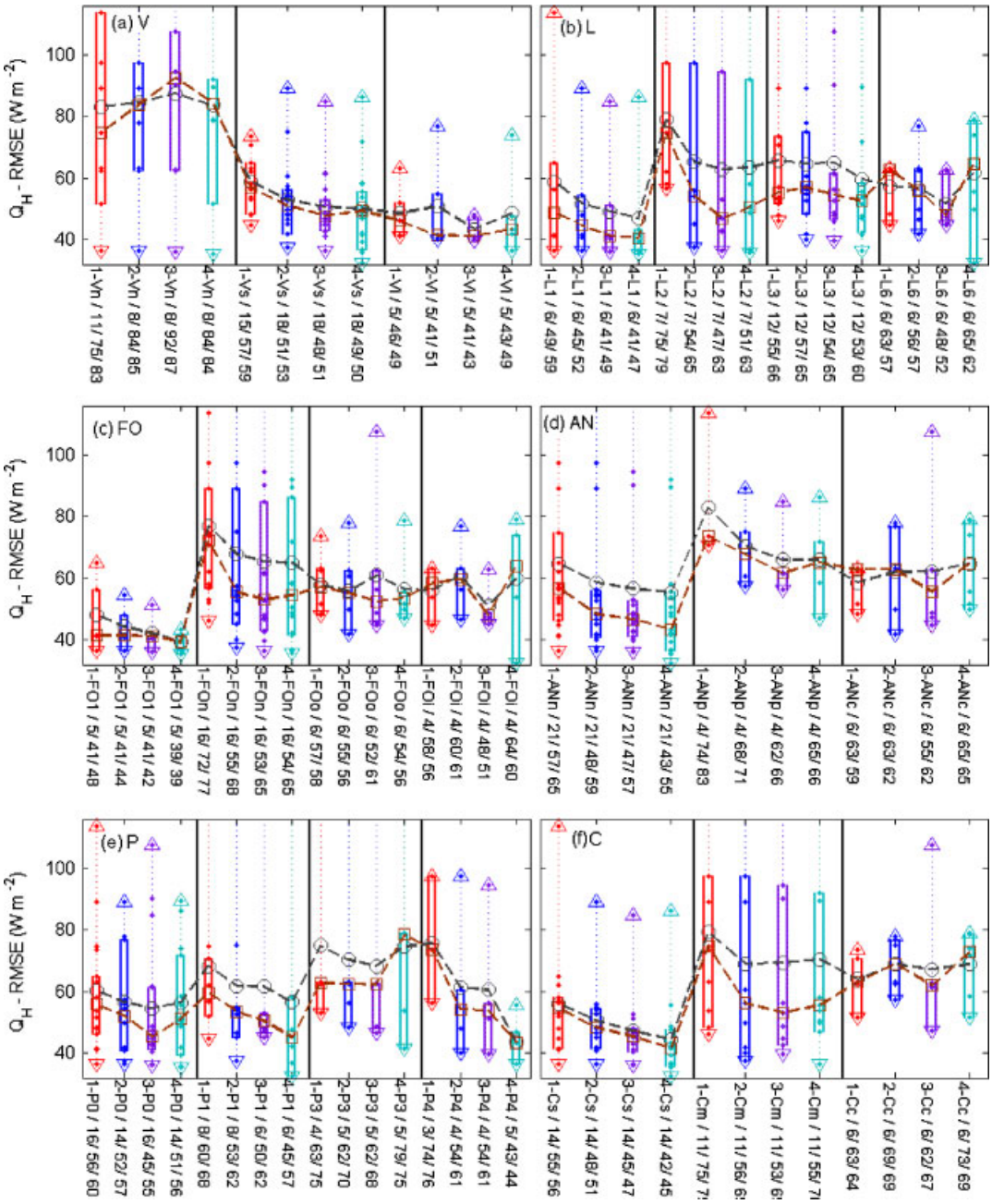

Figure 13. As for Figure 4 but for turbulent sensible heat flux (Q_H) for all hours. Note plots are cut-off at 0.90 of the maximum. This figure is available in colour online at wileyonlinelibrary.com/journal/joc

RMSE are small. The best ensemble performance is for the simple models, followed by the all ensemble. After Stage 1, the medium and complex models have a very similar ensemble performance.

From Stage 1 to 2, three more models chose to include vegetation (Figure 16(a)). The three models which incorporated vegetation did so by using separate vegetation tile(s) (Vs). The Vs approach, the most common, had a 10 W m^{-2} improvement between Stages 1 and 2. This is because in Stage 2 the separate tiles can be more realistically weighted. For Vs models, there is a reduction in the mean RMSE at each stage. For the Vs models, except after rainfall, the latent heat flux is coming exclusively from the vegetation scheme that has been 'coupled' to the urban scheme. These schemes have been extensively tested in earlier PILPS studies; however, they have not been extensively tested for use in urban areas. The user has to decide which vegetation type to select (see discussion in Grimmond *et al.*, 2010) as well as the appropriate parameter values for that vegetation class.

The simpler models which take a bulk approach to the urban morphology (L1) initially have the smallest median RMSE compared to more complex models (L2, L3, L6) (43/58/56/56 W m^{-2}) (Figure 16(b)). The L1 models do improve with subsequent stages but the range also becomes larger. The improvement, however, is not as great by Stage 4 as that which occurs for the L2/L3/L6 (39/38/45/48 W m^{-2}). The L2 models thus improve the

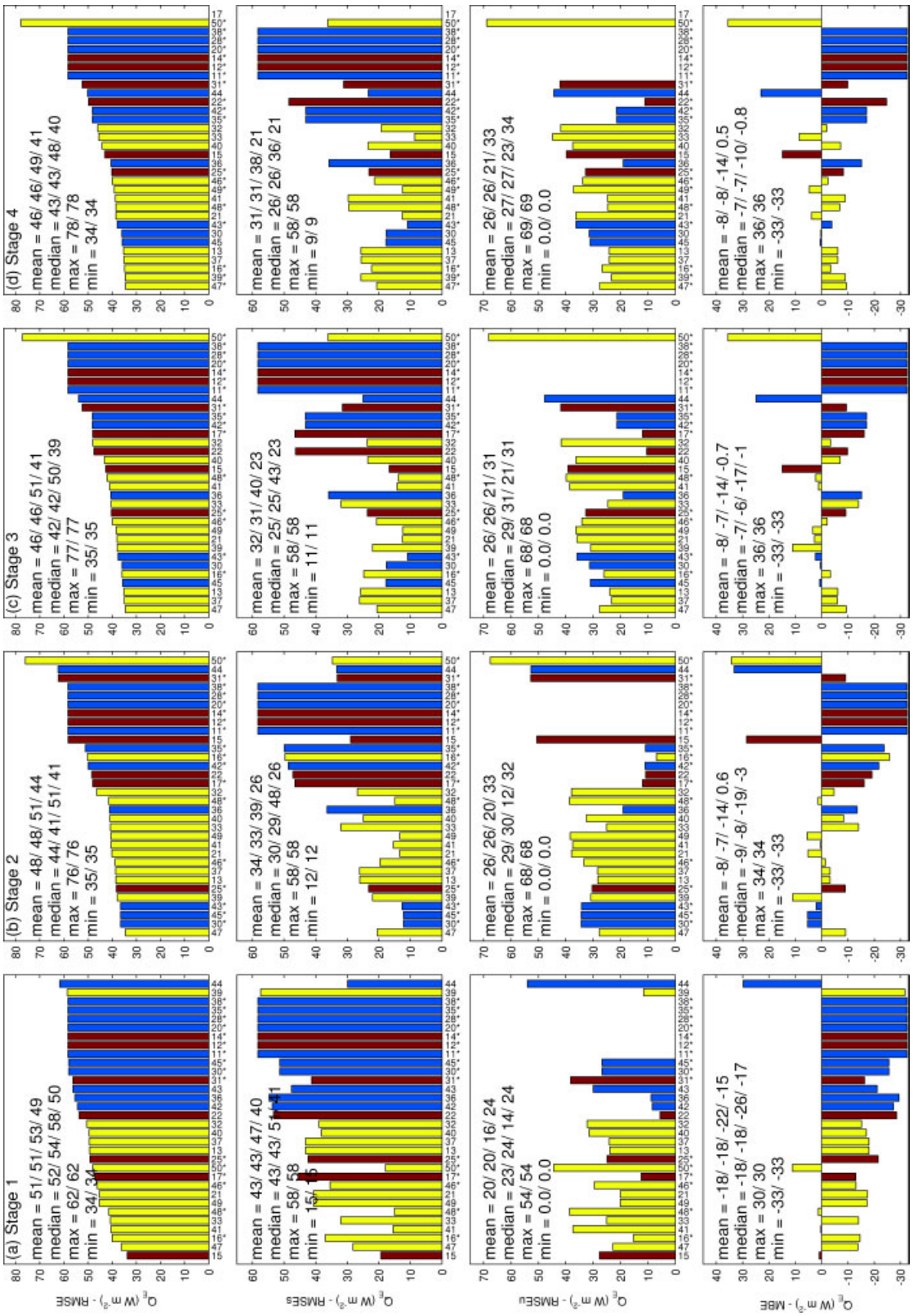

Figure 14. As for Figure 2 but for turbulent latent heat flux (Q_E) for all hours. The mean observed flux for this period was 32.5 W m^{-2}. This figure is available in colour online at wileyonlinelibrary.com/journal/joc

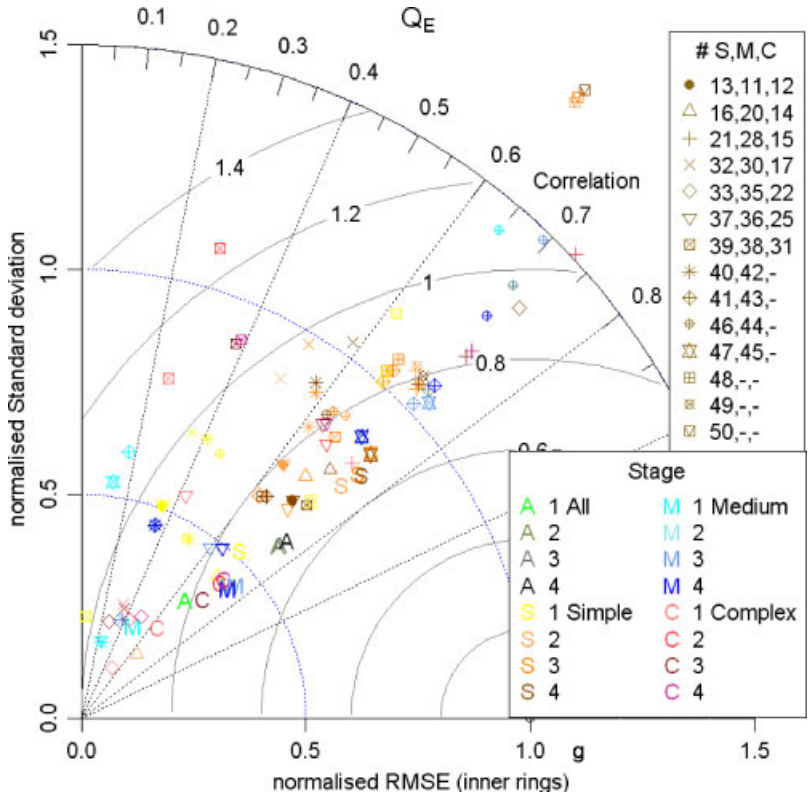

Figure 15. As for Figure 3 but for turbulent latent heat flux (Q_E) for all hours. This figure is available in colour online at wileyonlinelibrary.com/journal/joc

most and have the lowest median RMSE. A small improvement is seen in the median across all four facet and orientations classes (FO) by Stage 4. The models which do not distinguish facets (FO1) have the smallest median RMSE at Stage 4 but the greatest improvement is for those models that have facets but do not account for orientation (FOn).

The models with radiative closure (P0) have a larger median RMSE at Stage 4 than the P1 and P4 models. The P4 models have improvement at each of the four stages but have a slightly larger median RMSE at Stage 4 than the P1 models. The P3 models show no change in the median with stage as many do not model vegetation. Overall, the simplest models (Cs) perform best at all four stages but the Cm models have a greater gain from the additional information provided across the four stages (Figure 16(f)).

The daytime RMSE values are larger than for the night-time period (Stage 1 median 71/21 W m^{-2}) and all hours which is when the observed flux is larger and more variable (mean observed flux day = 56.41 W m^{-2}, night = 8.53 W m^{-2}). The night-time fluxes do not show any improvement in performance over the four stages and there is little variation between methods. At Stage 4, the daytime RMSE is 57 W m^{-2}. The simplest models (Cs) have a median RMSE that is the smallest with a RMSE of 51 W m^{-2} and have a 10 W m^{-2} improvement over the four stages.

The turbulent heat fluxes are not modelled as well as the radiative fluxes. But as with the radiative fluxes the inclusion of vegetation improves model performance. However, despite in Stage 4 knowing the site location, many models did a poorer job than at previous stages.

Overall, the simple models (Cs) do the best job of modelling latent heat flux. They also systematically improve as the additional information becomes available. Taking vegetation into account is critical to model Q_E appropriately. The models that use the separate tile scheme have about the same overall performance as those that take an integrated approach. But there is a much wider range of results from the separate tile models. This suggests that using vegetation schemes that have been tested in nonurban areas are better than ignoring vegetation, but given the wide range of results it suggests that some careful thought may need to be given to ensure their use is appropriate. Here we have not investigated whether the modellers assumed any additional water, such as irrigation, to be available for evaporation.

4. Conclusions

Groups around the world have run ULSMs in offline mode for four stages, with increasing information about the site provided. Initially, the groups knew only that the site was urban but by Stage 4 detailed surface materials characteristics had been provided. Here the ability to model the radiation and energy balance fluxes on average for a year is evaluated. It should be remembered that observations also have errors which vary with time of day, season, latitude, local geography, and land cover.

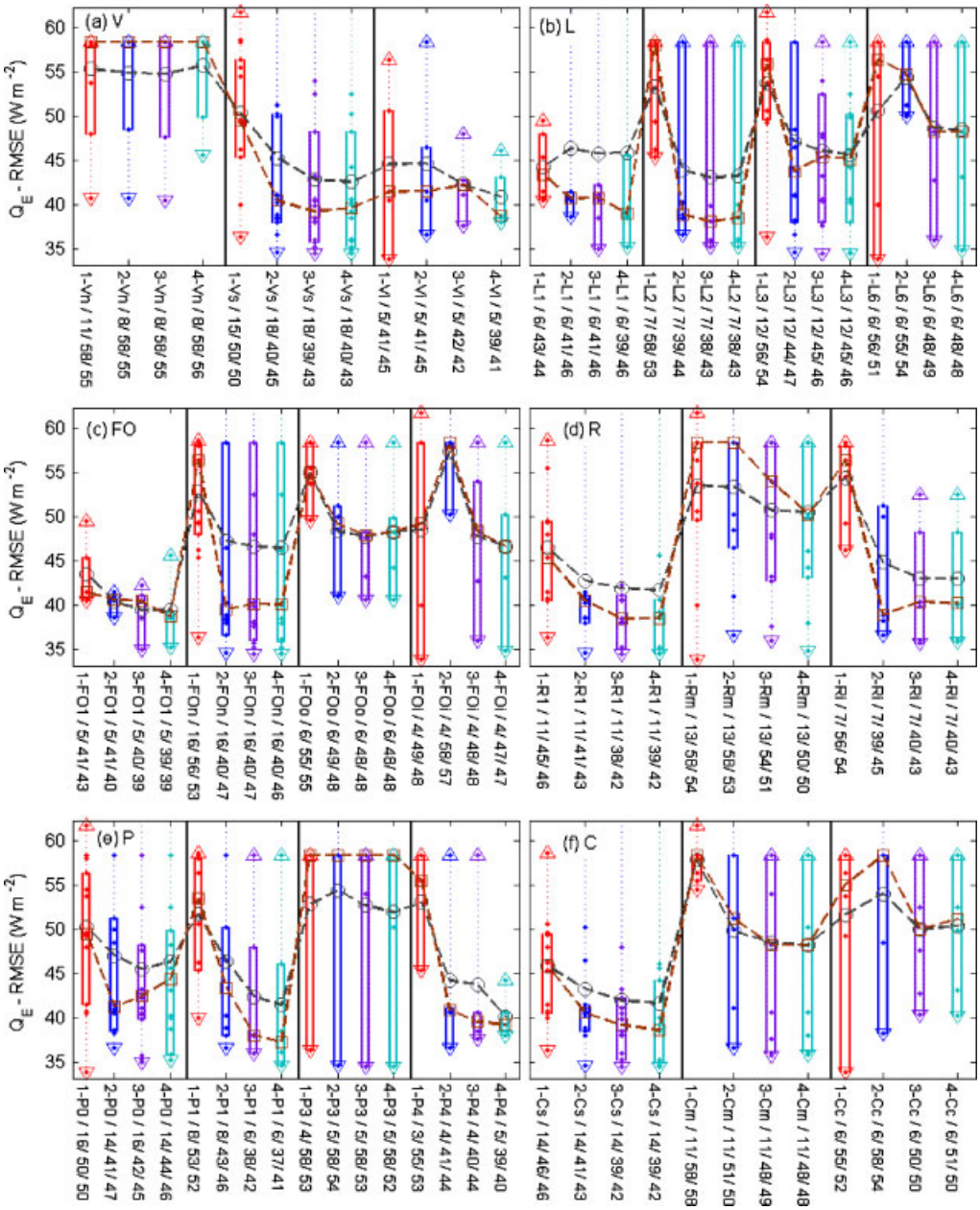

Figure 16. As for Figure 4 but for turbulent latent heat flux (Q_E) for all hours. Note plots are cut-off at 0.80 of the maximum. This figure is available in colour online at wileyonlinelibrary.com/journal/joc

In the process of running the models, a small number of models improved for an individual flux at each stage as new information was provided (2, 2, 2, 5, 4; for $K_\uparrow$, $L_\uparrow$, Q^*, Q_H, Q_E, respectively). However, there are other models that have a drop in performance with the addition of more complete information, and cases where there is a systematic decline at all stages (0, 1, 4, 0, 1; for $K_\uparrow$, $L_\uparrow$, Q^*, Q_H, Q_E, respectively).

From the analysis of the data returned from the modelling groups in relation to the observed flux data the following conclusions are drawn:

- A wide range of model performance is evident for each flux. No individual model does best for every flux modelled. Clearly this finding has very significant implications for the application of any model. It may also imply that in some cases models perform well but for the wrong physical reasons. For example, if a model overestimates the net shortwave radiation, but accurately models the sensible heat flux, then it may indicate a problem also in the physical representation of the heat exchanges between the surfaces and the atmosphere (since it needs to 'absorb' more energy to get the right sensible heat flux).
- Taking vegetation cover into account (or not) significantly impacts model performance. This conclusion is in agreement with those of Phase 1 (Grimmond *et al.*, 2010) where the site had a much lower plan area fraction of vegetation than the Phase 2 site. Data provided at Stage 2 (surface cover fractions) usually had the

largest impact on model performance. Moreover, the fact that the RMSE for the latent heat flux is of the same order as the latent heat flux itself, indicates that work needs to be done to improve simulations of this flux.

- Closure of the radiation balance is not a good measure of the ability to calculate a particular radiative flux. Comparing the performance of the components of the radiation balance to the net all wave radiation shows a clear re-ranking between fluxes. Notably those that perform poorly for an individual component flux are not the poorest for Q^* (Figures 2, 5, and 8). This means it is important that when a user applies a model they are aware of the performance of the ULSM not only for the initial flux of interest but also for the other fluxes for which the user may wish to infer impact. Given the increasing use of ULSM for assessing mitigation and adaptation strategies this is very important.
- Overall, the ULSM generally model $K_{\uparrow}$ well and additional surface information does result in an improvement of performance. The models are able to estimate reasonably well the amount of energy absorbed by the urban fabric, but have bigger problems in partitioning it between longwave, sensible, latent, and storage heat fluxes.
- Overall $L_{\uparrow}$ is not as well modelled as $K_{\uparrow}$. The set of model characteristics that minimize the errors in the outgoing longwave radiation change with the time of day. Generally performance improved when the pervious/impervious fraction became known but did not when heights and further information on surface fractions were provided. The performance of most models deteriorated when building material information was provided; typically back to the levels at Stage 1 but in many cases even poorer. Given the difficulty to gather appropriate values of material characteristics, their provision may not currently be worth the effort given how models then perform. Alternatively, there is a need to ensure that the data are of much better quality than is currently 'easily' obtainable.
- Net all wave radiation is modelled better than either $K_{\uparrow}$ or $L_{\uparrow}$. In general, the radiative fluxes are modelled better than the turbulent fluxes. The net all wave radiation is clearly the best modelled flux which is in agreement with Phase 1 results (Grimmond *et al.*, 2010). There is clear trade-off in performance between net all wave radiation (Q^*) and turbulent sensible heat flux (Q_H) which is in agreement with Loridan *et al.* (2010a).
- The errors from the models were smaller during the night than they were during the daytime, although this might be expected as the surface energy balance is not dominated by the solar radiation during this period.
- For the net radiation, simple characteristics (L1/FO1/R1/AE1/Cs) give the best results for both daytime and night-time, although there is much greater variability between the classes than for the outgoing longwave radiation.
- The models that perform best, for individual characteristics, are those that are the simplest as they can be assigned one parameter that is close to the observed value. Based on overall complexity the simplest and the most complex have similar results which are better than the medium complexity models.
- Additional surface information is important in improving model performance. However, there is evidence that good model physics is not enough to prevent the users' choice of parameter values from significantly influencing the outcome. Therefore, it is essential when models are being used for scenario testing that appropriate parameter values are used.
- Simpler models often showed a net improvement with additional information; the more complex models did not. This may be because there was not enough additional detailed information provided so it was more difficult for the users to decide how to use this information appropriately. It is important to note that parameters specified for simpler models (e.g. overall albedo) often equate to empirical aggregations of processes in more complex models (e.g. the net effect of reflections due to facet albedos). Nevertheless, the results here suggest that increased model complexity does not necessarily increase model performance.
- It is expected that more complex models may have more potential for future improvements as they are able to resolve more details without deteriorating their performance. The most complex models are more flexible and have the potential to describe the biophysical interactions between the atmosphere and urban surfaces. Although the ability to do this has not been tested here, these models can provide vertical profiles of atmospheric variables within the urban canopy layer. If the simulation is for weather forecasting, a good estimate of the heat fluxes at the top of the urban canopy is probably sufficient, and, consequently, a simple scheme may be the appropriate choice. If air quality is the focus, the atmospheric behaviour within the urban canopy layer may be important, and a more complex scheme can be useful. An important finding of PILPS-urban is that in many cases, work is needed to improve the complex schemes (both in terms of physics and definition of numerical constants), in order to have skills comparable to those of the more simple schemes in estimating energy fluxes at the top of the urban canopy. More complicated models are generally more difficult to use and it is even difficult for modellers to identify which are the most critical points of their model.
- As a community it is clear that in terms of surface characteristics, the information up to Stage 3 (Table II) benefited a large number of models. The AE were also beneficial (Stage 4) but the provision and acquisition of the most appropriate wall, roof, and road thermal properties need further thought and development from the modelling community. This model intercomparison

has already generated a suggested improved method for thermal parameter specification that accounts for the high degree of heterogeneity of such parameters in cities (Salamanca *et al.* 2009, 2010). Hopefully, additional analyses will shed more light on this issue.
- Using an ensemble of models rather than one model is generally better than any individual model for an individual flux. In general, the medium complexity ensemble performs least well and the simple performs best. The 'all' ensemble is always better than the medium complexity. Given the overall better performance of the ensembles they may be better than using one individual model when considering all of the fluxes.

These results are the first of a number of different studies that will be undertaken from these model runs. Future analyses will consider the role of seasonality on model performance, role of cloud conditions (day and night), time since rainfall, wind regime, the range of parameter values that are used, and the determination of optimized parameters; the participants will also analyse what they have learnt from the model comparison. To date, only two urban sites have been compared (Phase 1 and 2), which obviously is not representative of the wide range of land covers and morphologies, etc. found within neighbourhoods around the globe. However, some common conclusions are arrived at from comparison with these two sites, such as the best ability is for modelling net all wave radiation flux. Most notably, despite the range of vegetation cover found at the two sites, accounting for vegetation appears to be essential when modelling urban surface energy flux exchanges. There is a need for future comparisons of this type for sites with varying morphology and across a wider range of building materials. Our initial message is one of caution in applying any ULSM because, in general, no model performs well across all fluxes and it may be best to use an ensemble approach.

Acknowledgements

We would like to thank the IAUC/ICUC 7 editors of this volume. Funds to support this work (Grimmond) have included: Met Office (P001550), EU framework 7 (7 FP7-ENV-2007-1) projects MEGAPOLI (212520), BRIDGE (211345), and NSF (ATM-0710631). This work contributes to COST728. We would like to thank all who were involved in the collection of the original dataset and organisations and individuals who provided data that supported the work. The following people provided information about the site characteristics: Faculty of Architecture, Building and Planning, University of Melbourne (David O'Brian, John Sadar, Hamish Hill, Jon Robinson, Anna Hurlimann, Julie Willis, Valerie Francis, Carolyn Whitzman), Centre for Sustainable Architecture with Wood, School of Architecture & Design, University of Tasmania (Gregory Nolan), City of Darebin Council (David Archer), TPC Solutions Pty. Ltd. (Boris Iskra), Forest and Wood Products Australia, Ltd. (Jugo Ilic), Centre for Sustainable Infrastructure, Civil Engineering, Swinburne University of Technology (Kerry McManus). Funding from CATER 2006-2202 (Baik), BSIK-COM29 (Steeneveld), and all the agencies that support the considerable time contributed by participating groups are acknowledged.

References

Best MJ. 2005. Representing urban areas within operational numerical weather prediction models. *Boundary-Layer Meteorology* **114**: 91–109.

Best MJ, Grimmond CSB, Villani MG. 2006. Evaluation of the urban tile in MOSES using surface energy balance observations. *Boundary-Layer Meteorology* **118**: 503–525.

Chen F, Kusaka H, Tewari M, Bao J, Hirakuchi H. 2004. Utilizing the coupled WRF/LSM/Urban modeling system with detailed urban classification to simulate the urban heat island phenomena over the Greater Houston area. *Fifth Symposium on the Urban Environment*, CD-ROM. 9.11. Amer. Meteor. Soc., Vancouver, BC, Canada.

Clarke JA, Yaneske PP, Pinney AA. 1991. *The Harmonisation of Thermal Properties of Building Materials*. BEPAC Publication: Watford, UK; TN91/6, ISBN 0 187 212 607 3.

Coutts AM, Beringer J, Tapper NJ. 2007a. Characteristics influencing the variability of urban CO2 fluxes in Melbourne, Australia. *Atmospheric Environment* **41**: 51–62.

Coutts AM, Beringer J, Tapper NJ. 2007b. Impact of increasing urban density on local climate: spatial and temporal variations in the surface energy balance in Melbourne, Australia. *Journal of Applied Meteorology* **47**: 477–493.

Dandou A, Tombrou M, Akylas E, Soulakellis N, Bossioli E. 2005. Development and evaluation of an urban parameterization scheme in the Penn State/NCAR Mesoscale model (MM5). *Journal of Geophysical Research* **110**: D10102.

Dragoni D, Schmid HP, Grimmond CSB, Loescher H. 2007. Uncertainty of annual net ecosystem productivity estimated using eddy-covariance flux measurements. *Journal of Geophysical Research* **112**: D17102, DOI:10.1029/2006JD008149.

Dupont S, Mestayer PG. 2006. Parameterisation of the urban energy budget with the submesoscale soil model. *Journal of Applied Meteorology and Climatology* **45**: 1744–1765.

Dupont S, Mestayer PG, Guilloteau E, Berthier E, Andrieu H. 2006. Parameterisation of the urban water budget with the submesoscale soil model. *Journal of Applied Meteorology and Climatology* **45**: 624–648.

Engineering Toolbox. 2005a. Gases – specific heat capacities and individual gas constants [Online]. Available from: http://www.engineeringtoolbox.com/spesific-heat-capacity-gases-d_159.html (Last accessed March 2010).

Engineering Toolbox. 2005a. Thermal conductivity of some common materials [Online]. Available from: http://www.engineeringtoolbox.com/thermal-conductivity-d_429.html (Last accessed March 2010).

Essery RLH, Best MJ, Betts RA, Cox PM, Taylor CM. 2003. Explicit representation of subgrid heterogeneity in a GCM land surface scheme. *Journal of Hydrometeorology* **4**: 530–543.

Foken T. 2008. *Micrometeorology*. Springer: Berlin, Heidelberg; 308 pp, ISBN: 978-3-540-74665-2.

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

Fortuniak K, Offerle B, Grimmond CSB. 2004. Slab surface energy balance scheme and its application to parameterisation of the energy fluxes on urban areas. NATO ASI, Kiev, Ukraine; 82–83. Available from: www.met.rdg.ac.uk/urb_met/NATO_ASI/talks.html (Last accessed 4–15 May 2010).

Fortuniak K, Offerle B, Grimmond CSB. 2005. Application of a slab surface energy balance model to determine surface parameters for urban areas. *Lund Electronic Reports in Physical Geography* **5**: 90–91.

Gillett NP, Zwiers FW, Weaver AJ, Hegerl GC, Allen MR, Stott PA. 2002. Detecting anthropogenic influence with a multi-model ensemble. *Geophysical Research Letters* **29**: 1970, DOI:10.1029/2002GL015836.

Grimmond CSB, Oke TR. 1991. An evaporation-interception model for urban areas. *Water Resources Research* **27**: 1739–1755.

Grimmond CSB, Oke TR. 1999. Heat storage in urban areas: observations and evaluation of a simple model. *Journal of Applied Meteorology* **38**: 922–940.

Grimmond CSB, Oke TR. 2002. Turbulent heat fluxes in urban areas: observations and local-scale urban meteorological parameterization scheme (LUMPS). *Journal of Applied Meteorology* **41**: 792–810.

Grimmond CSB, Best M, Barlow J, Arnfield AJ, Baik J-J, 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, Williamson T. 2009. Urban surface energy balance models: model characteristics and methodology for a comparison study. In *Meteorological and Air Quality Models for Urban Areas*, Baklanov A, Grimmond CSB, Mahura A, Athanassiadou M (eds). Springer-Verlag: Berlin, Heidelberg; ISBN: 978-3-642-00297-7.

Grimmond CSB, Blackett M, Best MJ, Barlow J, Baik J-J, Belcher SE, Bohnenstengel SI, Calmet I, Chen F, Dandou A, Fortuniak K, Gouvea ML, Hamdi R, Hendry M, Kawai T, Kawamoto Y, Kondo H, Krayenhoff ES, Lee S-H, Loridan T, Martilli A, Masson V, Miao S, Oleson K, Pigeon G, Porson A, Ryu Y-H, Salamanca F, Shashua-Bar L, Steeneveld G-J, Trombou M, Voogt J, Young D, Zhang N. 2010. The international urban energy balance models comparison project: first results from phase 1. *Journal of Applied Meteorology and Climatology* **49**: 1268–1292, DOI: 10.1175/2010JAMC2354.1.

Hamdi R, Schayes G. 2007. Validation of Martilli's urban boundary layer scheme with measurements from two mid-latitude European cities. *Atmospheric Chemistry and Physics* **7**: 4513–4526.

Hamdi R, Masson V. 2008. Inclusion of a drag approach in the Town Energy Balance (TEB) scheme: offline 1-D evaluation in a street canyon. *Journal of Applied Meteorology and Climatology* **47**: 2627–2644.

Harman IN, Best MJ, Belcher SE. 2004a. Radiative exchange in an urban street canyon. *Boundary-Layer Meteorology* **110**: 301–316.

Harman IN, Barlow JF, Belcher SE. 2004b. Scalar fluxes from urban street canyons. Part II: model. *Boundary-Layer Meteorology* **113**: 387–410.

Harman IN, Belcher SE. 2006. The surface energy balance and boundary layer over urban street canyons. *Quarterly Journal of the Royal Meteorological Society* **132**: 2749–2768.

Henderson-Sellers A, Yang ZL, Dickenson RE. 1993. The project for intercomparison of land-surface parameterization schemes. *Bulletin of the American Meteorological Society* **74**: 1335–1349.

Henderson-Sellers A, Irannejad P, McGuffie K, Pitman A. 2003. Predicting land-surface climates-better skill or moving targets?. *Geophysical Research Letters* **30**: 1777, DOI:10.1029/2003GL017387.

Hollinger DY, Richardson AD. 2005. Uncertainty in eddy covariance measurements and its application to physiological models. *Tree Physiology* **25**: 873–885.

Irranejad P, Henderson-Sellers A, Sharmeen S. 2003. Importance of land-surface parameterization for latent heat simulation in global atmospheric models. *Geophysical Research Letters* **30**: 1904, DOI:10.1029/2003/GL018044.

Jacobson MZ. 1999. *Fundamentals of Atmospheric Modeling*. Cambridge University Press: Cambridge.

Kanda M, Kawai T, Kanega M, Moriwaki R, Narita K, Hagishima A. 2005a. A simple energy balance model for regular building arrays. *Boundary-Layer Meteorology* **116**: 423–443.

Kanda M, Kawai T, Nakagawa K. 2005b. A simple theoretical radiation scheme for regular building arrays. *Boundary-Layer Meteorology* **114**: 71–90.

Kawai T, Kanda M, Narita K, Hagishima A. 2007. Validation of a numerical model for urban energy-exchange using outdoor scale-model measurements. *International Journal of Climatology* **27**: 1931–1942.

Kawai T, Ridwan MK, Kanda M. 2009. Evaluation of the simple urban energy balance model using 1-yr flux observations at two cities. *Journal of Applied Meteorology and Climatology* **48**: 693–715.

Kawamoto Y, Ooka R. 2006. Analysis of the radiation field at pedestrian level using a meso-scale meteorological model incorporating the urban canopy model. In *ICUC-6*, Göteborg, Sweden, 12–16 June 2006.

Kawamoto Y, Ooka R. 2009a. Accuracy validation of urban climate analysis model using MM5 incorporating a multi-layer urban canopy model. In *ICUC-7*, Yokohama, Japan, 28 June–3 July 2009.

Kawamoto Y, Ooka R. 2009b. Development of urban climate analysis model using MM5 Part 2 – incorporating an urban canopy model to represent the effect of buildings. *Journal of Environmental Engineering (Transactions of AIJ)* **74**(642): 1009–1018 (in Japanese).

Kondo H, Liu FH. 1998. A study on the urban thermal environment obtained through a one-dimensional urban canopy model. *Journal of Japan Society for Atmospheric Environment* **33**: 179–192 (in Japanese).

Kondo H, Genchi Y, Kikegawa Y, Ohashi Y, Yoshikado H, Komiyama H. 2005. Development of a multi-layer urban canopy model for the analysis of energy consumption in a big city: structure of the urban canopy model and its basic performance. *Boundary-Layer Meteorology* **116**: 395–421.

Krayenhoff ES, Voogt JA. 2007. A microscale three-dimensional urban energy balance model for studying surface temperatures. *Boundary-Layer Meteorology* **123**: 433–461.

Kusaka H, Kondo H, Kikegawa Y, Kimura F. 2001. A simple single-layer urban canopy model for atmospheric models: comparison with multi-layer and slab models. *Boundary-Layer Meteorology* **101**: 329–358.

Lee S-H, Park S-U. 2008. A vegetated urban canopy model for meteorological and environmental modelling. *Boundary-Layer Meteorology* **126**: 73–102.

Lee X, Massman WJ, Law BE. 2004. *Handbook of Micrometeorology*, Kluwer Academic Publishers: Dordrecht.

Lemonsu A, Grimmond CSB, Masson V. 2004. Modelling the surface energy balance of an old Mediterranean city core. *Journal of Applied Meteorology* **43**: 312–327.

Loridan T, Grimmond CSB, Grossman-Clarke S, Chen F, Tewari M, Manning K, Martilli A, Kusaka H, Best M. 2010a. Trade-offs and responsiveness of the single-layer urban canopy parameterization in WRF: an offline evaluation using the MOSCEM optimization algorithm and field observations. *Quarterly Journal of the Royal Meteorological Society* **136**: 997–1019, DOI:10.1002/qj.614.

Loridan T, Grimmond CSB, Offerle BD, Young DT, Smith T, Jarvi L. Lindberg 2010 Local-Scale Urban Meteorological Parameterization Scheme (LUMPS): longwave radiation parameterization & seasonality related developments *Journal of Applied Meteorology & Climatology*, Doi: 10.1175/2010JAMC2474.1.

Martilli A, Clappier A, Rotach MW. 2002. An urban surface exchange parameterisation for mesoscale models. *Boundary Layer Meteorology* **104**: 261–304.

Masson V. 2000. A physically-based scheme for the urban energy budget in atmospheric models. *Boundary-Layer Meteorology* **41**: 1011–1026.

Masson V, Grimmond CSB, Oke TR. 2002. Evaluation of the Town Energy Balance (TEB) scheme with direct measurements from dry districts in two cities. *Journal of Applied Meteorology* **41**: 1011–1026.

Masson V, Seity Y. 2009. Including atmospheric layers in vegetation and urban offline surface schemes. *Journal of Applied Meteorology and Climatology* **48**: 1377–1397.

Ochsner TE, Horton R, Renb T. 2001. A new perspective on soil thermal properties. *Soil Science Society of America Journal* **65**: 1641–1647.

Offerle B, Grimmond CSB, Oke TR. 2003. Parameterization of net all-wave radiation for urban areas. *Journal of Applied Meteorology* **42**: 1157–1173.

Offerle B, Grimmond CSB, Fortuniak K. 2005. Heat storage and anthropogenic heat flux in relation to the energy balance of a central European city center. *International Journal of Climatology* **25**: 1405–1419.

Oleson KW, Bonan GB, Feddema J, Vertenstein M, Grimmond CSB. 2008a. An urban parameterization for a global climate model: 1. Formulation and evaluation for two cities. *Journal of Applied Meteorology and Climatology* **47**: 1038–1060.

Oleson KW, Bonan GB, Feddema J, Vertenstein M. 2008b. An urban parameterization for a global climate model: 2. Sensitivity to input parameters and the simulated heat island in offline simulations. *Journal of Applied Meteorology and Climatology* **47**: 1061–1076.

Pigeon G, Moscicki MA, Voogt JA, Masson V. 2008. Simulation of fall and winter surface energy balance over a dense urban area using the TEB scheme. *Meteorology and Atmospheric Physics* **102**: 159–171.

Porson A, Clark PA, Harman IN, Best MJ, Belcher SE. 2010. Implementation of a new urban energy budget scheme in the MetUM. Part I: description and idealized simulations. *Quarterly Journal of the Royal Meteorological Society*, DOI:10. 1002/qj.668.

Porson A, Harman IN, Bohnenstengel SI, Belcher SE. 2009. How many facets are needed to represent the surface energy balance of an urban area? *Boundary-Layer Meteorology* **132**: 107–128.

Richardson AD, Hollinger DY, Burba GG, Davis K,J, Flanagan LB, Katul GG, Munger JW, Ricciuto DM, Stoy PC, Suyker AE, Verma SB, Wofsy SC. 2006. A multi-site analysis of random error in tower-based measurements of carbon and energy fluxes. *Agricultural and Forest Meteorology* **136**: 1–18, DOI:10.1016/j.agrformet.2006.01.007.

Roberts SM, Oke TR, Grimmond CSB, Voogt JA. 2006. Comparison of four methods to estimate urban heat storage. *Journal of Applied Meteorology and Climatology* **45**: 1766–1781.

Ryu Y-H, Baik J-J, Lee S-H. 2009. A new single-layer urban canopy model for use in mesoscale atmospheric models. In *Proceedings of the Seventh International Conference on Urban Climate, ICUC-7*, Yokohama, Japan, June 28–July 3 2009.

Sailor DJ, Lu L. 2004. A top-down methodology for developing diurnal and seasonal anthropogenic heating profiles for urban areas. *Atmospheric Environment* **38**: 2737–2748.

Salamanca F, Krayenhoff ES, Martilli A. 2009. On the derivation of material thermal properties representative of heterogeneous urban neighbourhoods. *Journal of Applied Meteorology and Climatology* **48**: 1725–1732.

Salamanca F, Krpo A, Martilli A, Clappier A. 2010. A new building energy model coupled with an urban canopy parameterization for urban climate simulations – part I. Formulation, verification, and sensitivity analysis of the model. *Theoretical and Applied Climatology*, DOI: 10.1007/s00704-009-0142-9.

Salamanca F, Martilli A. 2010. A new Building Energy Model coupled with an Urban Canopy Parameterization for urban climate simulations – part II. Validation with one dimension off-line simulations. *Theoretical and Applied Climatology* **99**: 345–356.

Taylor KE. 2001. Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research* **106**: 7183–7192.

Willmott CJ. 1981. On the validation of models. *Physical Geography* **2**: 184–194.

