MEGAPOLI Scientific Report 10-07
Urban Energy Balance Models Comparison
MEGAPOLI Deliverable 2.3
CSB Grimmond, M Blackett, MJ Best


London, 2010
Colophon

Serial title: MEGAPOLI Project Scientific Report 10-07

Title: Urban Energy Balance Models Comparison

Subtitle: MEGAPOLI Deliverable 2.3

Main Author(s): C.S.B. Grimmond\textsuperscript{1\textdagger}, M. Blackett\textsuperscript{1}, M.J. Best\textsuperscript{2}

Contributing Author(s): and J. Barlow\textsuperscript{3-J. Baik\textsuperscript{4}, S.E. Belcher\textsuperscript{3}, S.I. Bohnenstengel\textsuperscript{1}, I. Calmet\textsuperscript{5}, F. Chen\textsuperscript{6}, A. Dandou\textsuperscript{7}, K. Fortuniak\textsuperscript{3}, M.L. Gouvea\textsuperscript{1}, R. Hamdi\textsuperscript{9}, M. Hendry\textsuperscript{2}, T. Kawai\textsuperscript{10}, Y. Kawamoto\textsuperscript{11}, H. Kondo\textsuperscript{12}, E.S. Krayenhoff\textsuperscript{13}, S.-H. Lee\textsuperscript{4}, T. Loridan\textsuperscript{1}, A. Martilli\textsuperscript{14}, V. Masson\textsuperscript{15}, S. Miao\textsuperscript{16}, K. Oleson\textsuperscript{6}, G. Pigeon\textsuperscript{15}, A. Porson\textsuperscript{3-23}, Y.-H. Ryu\textsuperscript{4}, F. Salamanca\textsuperscript{14}, L. Shashua-Bar\textsuperscript{17}, G.-J. Steeneveld\textsuperscript{18}, M. Tombrou\textsuperscript{1}, J. Voogt\textsuperscript{19}, D. Young\textsuperscript{1}, N. Zhang\textsuperscript{20}, and J. Beringer\textsuperscript{21}, A Coutts\textsuperscript{21}, M Kanda\textsuperscript{22}, R Ooka\textsuperscript{11}

Responsible institution(s): 1King’s College London, UK, 2Met Office, UK, 3University of Reading, UK, 4Seoul National University, Korea, 5Laboratoire de Mécanique des Fluides UMR 6598 CNRS-Ecole Centrale de Nantes, France, 6National Center for Atmospheric Research, USA, 7National and Kapodistrian University of Athens, Greece, 8University of Lódz, Poland, 9Royal Meteorological Institute, Belgium, 10Ehime University, Japan, 11The University of Tokyo, Japan, 12National Institute of Advanced Industrial Science and Technology, Japan, 13University of British Columbia, Canada, 14CIEMAT, Spain, 15CNRM-GAME (URA 1357), Meteo France-CNRS, France, 16IUM, CMA, China, 17Ben Gurion University of the Negev, Israel, 18Wageningen University, Netherlands, 19University of Western Ontario, Canada, 20Nanjing University, China, 21Monash University, Australia, 22Tokyo Institute of Technology, Japan

\textdagger Corresponding author address:

Environmental Monitoring and Modelling Group
Department of Geography, King’s College London
London, WC2R 2LS, UK
E-mail: Sue.Grimmond@kcl.ac.uk

Language: English

Keywords: Urban climate, energy balance, surface atmosphere exchanges, land surface modelling, sustainable cities, radiation, turbulent heat fluxes, PILPS-urban

Url: http://megapoli.dmi.dk/publ/MEGAPOLI_sr10-07.pdf

Digital ISBN: 978-87-993898-0-3

MEGAPOLI: MEGAPOLI-10-REP-2010-03

Website: www.megapoli.info

Copyright:

# Content:

Abstract ................................................................................................................................................4

**Part I** ................................................................................................................................................5

**Parameterization of Urban Areas in Meso-Scale Models** .................................................................5

1. Introduction ..................................................................................................................................5

2. Momentum and turbulence production/ destruction ..................................................................6

3. Radiation ......................................................................................................................................6

4. Heat storage ................................................................................................................................7

5. Anthropogenic heat flux .............................................................................................................8

6. Sensible and latent heat flux .........................................................................................................8

Acknowledgements ............................................................................................................................9

References ..........................................................................................................................................9

**Part II** ............................................................................................................................................12

**International Urban Energy Balance Models Comparison Project: First Results from Phase 1** ..............................................................................................................................................42

1. Introduction ...................................................................................................................................12

2. The Characteristics of Urban Energy Balance Models ..............................................................16

2.1 Vegetation and latent heat flux .................................................................................................16

2.2 Anthropogenic heat fluxes .........................................................................................................16

2.3 Anthropogenic heat fluxes: temporal variation ........................................................................17

2.4 Urban morphology ......................................................................................................................17

2.5 Urban morphology: facets and orientation .............................................................................18

2.6 Radiative fluxes: reflections .........................................................................................................18

2.7 Radiative fluxes: albedo and emissivity ....................................................................................18

2.8 Storage heat flux .........................................................................................................................18

2.9 Other features ............................................................................................................................19

2.10 Turbulent sensible and latent heat fluxes .................................................................................19

2.11 Energy balance closure .............................................................................................................19

2.12 Anthropogenic water flux and other capabilities ....................................................................22

2.13 Model uniqueness ......................................................................................................................22

3. Model Inputs ................................................................................................................................23

4. The International Urban Energy Balance Model Project ...........................................................25

5. Results from VL92 .........................................................................................................................28

6. Conclusions ..................................................................................................................................36

Acknowledgements ............................................................................................................................37

References ..........................................................................................................................................37

**Part III** ...........................................................................................................................................42

**Initial Results from Phase 2 of International Urban Energy Balance Comparison Project** ...........................................................................................................................................42

1. Introduction ..................................................................................................................................42

2. Methodology ..................................................................................................................................43

3. Results ............................................................................................................................................48

3.1 Radiation comparison ..................................................................................................................48

3.2 Outgoing shortwave radiation .....................................................................................................49

3.3 Outgoing longwave radiation .....................................................................................................53

3.4 Net all wave radiation ..................................................................................................................56

3.5 Turbulent sensible heat flux ........................................................................................................59

3.6 Turbulent latent heat flux ..............................................................................................................61

References ..........................................................................................................................................66

Previous MEGAPOLI reports ................................................................................................................70
Abstract

A large number of urban surface energy balance models now exist with different assumptions about the important features of the surface and exchange processes that need to be incorporated. To date, no comparison of these models has been conducted; in contrast, models for natural surfaces have been compared extensively as part of the Project for Intercomparison of Land Surface Parameterisation Schemes. Here, the methods and first results from an extensive international comparison of 33 models are presented. The aim of the comparison overall is to understand the complexity required to model energy and water exchanges in urban areas. The degree of complexity included in the models is outlined and impacts on model performance are discussed. During the comparison there have been significant developments in the models with resulting improvements in performance (root mean square error falling by up to two-thirds). Evaluation is based on a dataset containing net all-wave radiation, sensible heat and latent heat flux observations for an industrial area in Vancouver. The aim of the comparison is two-fold: to identify those modelling approaches that minimise the errors in the simulated fluxes of the urban energy balance and to determine the degree of model complexity required for accurate simulations. There is evidence that some classes of models perform better for individual fluxes but no model performs best or worst for all fluxes. In general, the simpler models perform as well as the more complex models based on all statistical measures. Generally the schemes have best overall capability to model net all wave radiation and least capability to model latent heat flux.
Part I

Parameterization of Urban Areas in Meso-Scale Models

C.S.B. Grimmond

1. Introduction

In the last decade the number of models that simulate urban areas within meso-scale models has increased significantly (many of these are described and evaluated in Grimmond et al. 2009a,b). Globally, greater areal extent of urban areas and rapidly growing populations, a decrease in grid cell size in numerical models and enhanced computing capacity, make the representation of the urban land surface more frequently required.

Urban land surface schemes (ULSS) model the bulk, or area-average, response for the urban grid cell, not the response from individual roughness elements (buildings, trees etc) within the urban canopy layer. The lowest layer of the meso-scale model typically is approximately the ‘blending height’ which may be taken to be above the roughness sub-layer (i.e. a local scale or neighbourhood response of the surface), although there are exceptions to this (e.g. BEP model of Martilli et al. 2002).

To model the land surface a number of approaches have been taken, with knowledge of the urban surface present and how this has and is changing with time. Many models have prescribed surface classes with set parameter values. These may be appropriate, but in some cases it has been demonstrated their specification results in poorer model performance than if site specific values are assigned (Lemonsu et al. 2004).

In meso-scale models, “tiled” land surface schemes are often used (Essery et al. 2003). These require the fraction of surface type for each tile from which the weighted surface variable (flux) is calculated and passed to the next layer in the meso-scale model. Currently, in many ULSS, the tile only represents the built fraction and assumes that vegetation does not interact with the built fraction (or any other tile) until the first layer of the meso-scale model. As most land surface schemes include many vegetation types, it is possible to represent a range of vegetation types within the grid. Individual models set limits as to the number of tiles that can be used per grid cell. An alternative approach is to treat the surface as an integrated whole so that the ULSS incorporates the net sub-grid cell interactions between different surface types (e.g. advective effects between vegetation and built surfaces). Examples of this approach are SUES (Grimmond and Oke, 1991), LUMPS (Grimmond and Oke 2002) and CLMU (Oleson et al. 2008a).

A more complex approach is to have the urban classes with their characteristics (e.g. fraction built and spacing) while allowing for a distribution of building heights rather than one average height, within the cell (e.g. Martilli et al. 2002). This is especially useful when the models are used to drive dispersion or air pollution applications as the variability of roof height and shape has been demonstrated in wind tunnel studies to influence both roughness characteristics and surface exchange coefficients (Raffilidas, 1997; Barlow and Belcher 2002).

In terms of actual processes, the approaches taken to incorporating urban features vary both in terms of those that are considered and in the methods that are used to capture their dynamics. The net result is a wide range of approaches taken to model the momentum, heat and water exchanges (Grimmond et al. 2009b) in varying combinations in meso-scale models. It is absolutely critical that
any user is aware of the approach adopted by a given ULSS and the inherent limitations in the results generated.

2. Momentum and turbulence production/destruction

Three general approaches have been used to account for the influence of the urban environment on momentum fluxes. The simplest alters the roughness length for momentum and zero plane displacement ($z_{0M}, z_d$) to appropriate urban values. The wind profile is modified by the change in boundary conditions of the lowest grid. However, using the log law means that if a wind profile within the roughness elements is needed, another approach must be taken.

Alternatively, the drag approach uses a modification of the momentum equations, through the addition of the drag force to the conservation of momentum equation. The exchanges with the atmosphere occur at ground level and at the atmospheric levels in contact with the buildings. Models can represent profiles of turbulence statistics in the canyon and the roughness sub layer but these additional terms can make coupling an ULSS and a meso-scale model more complicated (Masson, 2006).

The attenuation approach uses an equation based on wind profile observations within the canopy (Brown 2000) and, as such, can provide a wind speed below the logarithmic profile, but this is constrained by requiring appropriate values of attenuation coefficients.

The approaches taken to account for building enhanced turbulence parallel those for momentum. Modification of the roughness length in the log-law equation provides for changes in friction velocity ($u^*$). The turbulence closure scheme and boundary condition specification influence the eddy diffusivity, Reynolds shear stress, and/or the velocity gradient at the lowest level, each of which have dependence on $u^*$ (Brown, 2000). As noted, this is only applicable for heights above the surface for which the log-law applies. Models that use $k-\varepsilon$ or other turbulence closure approach to specify the eddy diffusivity, can have a turbulent kinetic energy (TKE) production term added to the TKE equation to account for the urban-induced turbulence (Brown, 2000). In addition, empirical scaling approaches can be used to determine local $u^*$ (e.g. Rotach 1997, Roth 2000).

3. Radiation

Radiation fluxes in cities are impacted by the urban atmosphere and the urban surface. The physical and chemical properties of urban atmosphere impact both the shortwave and the longwave radiation receipt. To incorporate this, meso-scale model radiation transfer schemes have been adapted in the atmospheric layers above the ULSS. For example, Martilli et al. (2002) used Schayes (1982), for solar radiation with a specific absorption factor for aerosols, and the Sasamori (1968) scheme for longwave radiation takes into account water vapour and carbon dioxide concentrations.

A number of approaches have been taken by ULSS to simulate the urban surface and its impact on radiative processes. The simplest considers the albedo ($\alpha$) and emissivity ($\varepsilon$) of urban materials. The surface may be represented as flat, ignoring the distinct canyon geometry of urban areas, with the net geometric effect taken into account by using $\alpha$ and $\varepsilon$ values determined for urban geometries and materials. The next level of sophistication divides the surface into more than two types (roof, road) and assigns $\alpha$ and $\varepsilon$ values, but still treats the urban surface as flat. The simplest approach which explicitly considers canyon geometry has three distinct surface types: wall, roof and road; each with different $\alpha$ and $\varepsilon$ values. In this approach, the solar geometry includes flat
surfaces (roof and road) and a vertical surface (wall). The height:width (H:W) and plan area fractions of roof and road determine solar access. However, with no orientation of the canyons, shaded or un-shaded areas are not considered at any given time. The premise for this approach is that when integrating over 360° all canyon orientations are equally likely (Masson 2000). In this approach, the canyon is also assumed to be infinitely long so that intersections are not resolved (Masson, 2000). More realistic models however, do include sunlit and shaded walls (Oleson et al. 2008a, b).

More complex approaches again incorporate two canyons with specific orientations (typically north-south and east-west) with the same H:W and material properties (α, ε) (Lee and Park, 2008). This allows for different fractions to be in sun and shade and, therefore, changes the net all wave radiation. However, no complete view factor geometry for all points is included. With increasing computer capacity, ULSS can be made more spatially explicit by facet, and within facet, to account for complexity of view factor geometry. However, typically the complete view factor geometry is not included in the ULSS but it is included in some micro-scale models (e.g. Krayenhoff and Voogt, 2007).

Different approaches also taken to account for reflections within the urban canyon. The approaches adopted have includes modelling one, multiple or infinite reflections. Harman et al. (2004a) concluded that canyon models need to consider at least one reflection of radiation and that multiple reflections are desirable for full applicability. The exact solution (Sparrow and Cess 1970) is needed for accuracy, if the α of the surface material is greater than 0.2 or the ε is less than 0.8, and increases as H:W increases. The errors associated with neglecting absorption, emission and scattering of radiation by the air in the canyon volume can be of a similar magnitude to those of neglecting multiple reflections (Verseghy and Munro, 1989a, b).

4. Heat storage

The assumptions made to incorporate urban features for radiation have parallels for the heat storage term. The simplest involve specification of the surface properties (thermal admittance, heat conductivity etc) based on a change to a net urban value. Complexity is added by assigning values for different facets (roof, wall, road) and then by increasing the number of layers. The more sophisticated models therefore allow for surface temperatures for facets, or facets with orientation, to be determined as the net balance between the radiative, conductive and convective fluxes within the urban surface scheme (e.g. CLMU Oleson et al. 2008a).

A number of models determine the heat storage based on surface temperatures (e.g. TEB, Masson 2000, BEP Martilli et al. 2002). Differences exist in terms of the way the three-dimensional surface is decomposed. For example, it can be treated as four 1-d element assemblages: building roofs, walls, internal mass and road surface, weighted by paved and vegetated fractions. Each assemblage is a series of layers, the depth of which is determined by the average volume of each, per unit plan area. The storage heat flux for the entire volume is considered (Offerle et al., 2005) and the average internal layer temperature is estimated by combining the 1-dimensional conservation of heat equation with Fourier’s Law. Another example is driven by internal and external facet temperatures and the thermal properties of building construction materials (Roberts et al. 2006).

An alternative, simple approach, the objective hysteresis model (Grimmond et al. 1991, Grimmond and Oke 1999), is driven by net all wave radiation. This approach parameterizes some of
the urban features (heat fluxes weighted by fraction of surface types) without requiring the additional computer and data resources of multi-layer conductive flux models.

5. Anthropogenic heat flux

The additional source of energy attributable to human activities needs to be resolved in models parameterising urban regions. Anthropogenic heat is derived from mobile sources (traffic), internal sources (domestic and industrial heat sources, for example) and from metabolism. This can be a very important source of energy, particularly in the densest part of a city (Oke 1988, Grimmond 1992, Ichinose et al. 1999). Its significance is proportionally greater in winter in extratropical regions given reduced solar radiation. There are two issues that need to be resolved with incorporating this flux. The first, and simpler, is the temporal pattern which varies seasonally, daily and through the week as a consequence of human activities. As such it can, relatively easily, be approximated. The more problematic issue, however, is the spatial variability of the flux. Activities at both “work” and “home” need to be considered, so it not appropriate to assign a flux based purely on housing or population density.

Some ULSSs have neglected the term; others have added it directly to the sensible heat flux that is passed to the mesoscale model or used it to modify internal temperatures and canyon air temperature. One approach to deal with the internal fixed sources, and the metabolic energy, is to explicitly specify an internal temperature for buildings that is maintained (with or without a temporal cycle) which affects heat storage and external surface temperature. However, this does not allow for the fact that additional energy is lost from buildings as part of maintaining these temperatures (both for heating and cooling). This additional energy needs to be added at roof level and/or in the canyon depending on season and typical building design. The mobile sources are located in the urban canyons so the energy needs to be added directly into the canyon airspace (if resolved).

6. Sensible and latent heat flux

Typically surface resistance (or its inverse, conductance) schemes are used to model the two convective fluxes. Depending on urban morphology, these consist of either one or multiple resistance networks which account for the number of facets and layers resolved. Slab models have the simplest resistance network. If the canyon is resolved, the exchange functions that occur both within the canyon and from the roof need to be considered. Depending on the degree of detail at which the canyon is resolved, it may be necessary to consider different surface temperatures for walls with different orientations, or to consider only one wall. For example, Masson (2000) has one gradient from the wall to the air, whereas the Martilli et al. (2002) model has four walls with different temperatures considered (N-S, E-W), each with a different gradient to the air temperature within the canyon. Martilli et al. (2002) further incorporate different heights of buildings and this is repeated for each layer.

A wide range of resistance schemes are used (e.g. Rowley 1930, Clarke 1985, Zilitinkevich 1995, Guilloteau 1998, Harman et al. 2004b). To determine the resistance, the wind profile within/above the canyon, and atmospheric stability, may be taken into account. Exchange between the canopy air and building surfaces may be parameterized by a roughness length or distributed sources of heat approach (generally in conjunction with a distributed drag approach). The number
of temperatures (and surface moistures) resolved, which drive the gradients, vary both for the surface and the air (within the canyon), related to morphology and number of vertical layers in the model. Monin-Obukhov similarity is assumed by many, although, probably it is not applicable within the urban canyon (Roth 2000). However, there is a lack of well-tested alternatives.

As $Q_H$ is calculated typically using surface temperature to force the gradient, a balance is inherent in the solution of surface temperature between the outgoing long wave radiation, $Q_H$ and heat conduction. Depending on the objective, performance is often improved for one flux at the expense of another. Models that use a combination type method (e.g. Penman-Monteith) do not require a surface temperature to determine the $Q_H$, but still need a resistance network to account for the transport of heat away from the surface.

The approaches taken to model resistances, surface temperature and air temperature occur in many combinations (see Grimmond et al. 2009b). For the calculation of sensible heat, the most common combinations are: one resistance from a single surface temperature to a single air temperature; and one resistance per facet from one surface temperature to a single air temperature.

The determination of the latent heat flux is theoretically the same as for the turbulent sensible heat, but with moisture gradient and diffusivities considered. However, a number of approaches have been taken to incorporate the surface control exerted by vegetation. The simplest, but unrealistic, approach is to ignore evaporation, thereby restricting a model to dry and unvegetated situations. This is the state of a number of current schemes but has been demonstrated to negatively impact model performance (Grimmond et al 2009c). With the tile approach, with urban (unvegetated) and vegetation tiles within a grid box, the latent heat fluxes (and other fluxes) for the vegetated fraction are calculated using the vegetation classes of a rural environment. This approach has been documented to yield reasonable results (Grimmond et al 2009b).

**Acknowledgements**

Funds to support this work (Grimmond): UK Met Office (P001550), EU framework 7 (7 FP7-ENV-2007-1) projects MEGAPOLI (212520), BRIDGE (211345). This is a contribution COST-728. The work of Matthew Blackett and Catherine Souch are gratefully acknowledged. The research leading to these results has also received funding from the European Union’s Seventh Framework Programme FP/2007-2011 under grant agreement n°212520.

**References**


Grimmond, C. S. B. and Oke, T. R., 1999: Heat storage in urban areas: Local-scale observations and


Grimmond et al. 2010: Part 3 of this document


Part II

International Urban Energy Balance Models Comparison Project: First Results from Phase 1

C.S.B. Grimmond, M Blackett, MJ Best


1. Introduction

The world’s population has become increasingly urbanised: around 29% of the global population were urban dwellers in 1950, 47% by 2000, and it is predicted to rise to 69% by 2050 (UN, 2007). Thus increasing numbers of people are impacted by weather and climate in urban areas. There is a growing requirement for accurate weather forecasts and climate change information within cities and concurrent increases in computer capabilities allow greater spatial resolution within models. In combination, there is a greater proportion of the Earth’s surface being categorized as “urban” and there are a larger number of smaller grid boxes in atmospheric models in which urban areas need to be resolved.

The surface morphology (i.e. urban form) and presence of impervious building materials, sparseness of vegetation and anthropogenic heat, water and pollutant contributions each have a significant effect on the climate of urban regions which lead to phenomena such as the urban heat island. Thus, effects of the urban surface on the fluxes of heat, moisture and momentum need to be accounted for in the land surface schemes used within numerical models, although the complexity of these schemes has to be balanced with their computational requirements. A fundamental aim of urban energy balance models is to accurately predict fluxes at the local scale (10²-10⁴ m). Some calculate additional terms including within-canyon air temperatures and wind speed, and facet surface temperature. A facet is a surface of the urban geometry that can be characterised by a single temperature and surface energy balance, and that can interact thermodynamically with other facets (for example, a wall facet exchanging longwave radiation with the road facet (Fig. 1). The outputs from the model may be hourly or higher temporal resolution for the whole surface, or be facet/orientation-specific.

Models have been developed to incorporate urban features for different applications ranging from: global climate modelling (e.g. Oleson et al. 2008a, b); numerical weather prediction (e.g. Best 1998, 2005, Masson 2000, Chen et al. 2004, Harman and Belcher 2006, Liu et al. 2006); air quality forecasting (e.g. Martilli et al. 2003) and dispersion modelling (e.g. Hanna and Chang 1992, 1993); characterization of measurements (e.g. Krayenhoff and Voogt 2007); and water balance modelling (e.g. Grimmond et al. 1986, Grimmond and Oke 1991). Across these schemes a wide range of urban features are incorporated; the models have varying levels of complexity, and different fluxes modelled (Tables 1, 2, Fig. 1).

In this paper, the methodology and initial results from the first international comparison of a broad range of urban land surface schemes are presented. The requirements of a land surface model from the perspective of an atmospheric model are considered, i.e. surface fluxes of heat, moisture
and momentum. Thus, the fundamental requirement for the models to be included is that they simulate urban energy balance fluxes. The forcing data for the surface models are the same as that which would be provided by an atmospheric model, i.e. the incoming short- ($K \downarrow$) and long-wave fluxes ($L \downarrow$), air temperature, specific humidity and the wind components. From these the outgoing radiative fluxes ($K \uparrow$, $L \uparrow$), net all wave radiation ($Q^*$), turbulent sensible heat flux ($Q_H$), turbulent latent heat flux ($Q_E$) and net heat storage flux ($\Delta Q_S$) are modelled. In this context, the net heat storage includes the energy storage within the buildings, the road and underlying soil and, for some models, the air space within the street canyon (Grimmond and Oke, 1999a). In the urban environment it is also useful to consider the anthropogenic heat flux ($Q_F$) in the surface energy balance (Oke 1988):

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

Features such as: additional sources of energy ($Q_F$), presence of built and natural surfaces, the bluff body nature of the buildings and existence of urban canyons, combine to change energy partitioning in urban areas. Thus significant modification to rural land parameterisation schemes is needed. Whilst many urban models have been evaluated against observational datasets (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), with some models even using the same observations, these comparisons have not been conducted in a controlled manner that allows robust model inter-comparison. The objective here is to do just that; to undertake a staged and carefully controlled classification and comparison of urban energy balance models and their performance. An important objective also is to determine which approaches minimise the errors in simulated fluxes.

<table>
<thead>
<tr>
<th>Code</th>
<th>Model Name</th>
<th>References</th>
<th>Versions</th>
<th>Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEP02</td>
<td>Building Effect Parameterization</td>
<td>Martilli et al. (2002)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>BEP_BEM08</td>
<td>BEP coupled with Building Energy Model</td>
<td>Martilli et al. (2002), Salamanca et al. (2009), Salamanca and Martilli (2009)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CLMU</td>
<td>Community Land Model - Urban</td>
<td>Oleson et al. (2008a, b)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GCTTC</td>
<td>Green Cluster Thermal Time Constant model</td>
<td>Shashua-Bar and Hoffman (2002; 2004)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>IISUCM</td>
<td>Institute of Industrial Science Urban Canopy Model</td>
<td>Kawamoto and Ooka (2006; 2009a; b)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>JULES</td>
<td>Joint UK Land Environment Simulator</td>
<td>Essery et al. (2003), Best (2005), Best et al. (2006)</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>LUMPS</td>
<td>Local-scale Urban Meteorological Parameterization Scheme</td>
<td>Grimmond and Oke (2002), Offerle et al. (2003)</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>NKUA</td>
<td>University of Athens Model</td>
<td>Dandou et al. (2005)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>MORUSES</td>
<td>Met Office Reading Urban Surface Exchange Scheme</td>
<td>Harman et al. (2004 a,b), Porson et al. (2009)</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>MUCM</td>
<td>Multi-layer Urban Canopy Model</td>
<td>Kondo and Liu (1998), Kondo et al. (2005)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NJU-UCM-S</td>
<td>Nanjing University Urban Canopy Model-single layer</td>
<td>Masson(2000), Kusaka (2001)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NJUC-UM-M</td>
<td>Nanjing University Urban Canopy Model-multiple layer</td>
<td>Kondo et al.(2005), Kanda(2005a; b)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SM2U</td>
<td>Soil Model for Submesoscales (Urbanized)</td>
<td>Dupont and Mestayer (2006), Dupont et al. (2006)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SNUUCM</td>
<td>Seoul National University Urban Canopy Model</td>
<td>Ryu et al. (2009)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SRUM2/SRUM4</td>
<td>Single Column Reading Urban Model tile version</td>
<td>Harman and Belcher (2006), Porson et al. (2009)</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>SUEB</td>
<td>Slab Urban Energy Balance Model</td>
<td>Fortuniak et al. (2004, 2005)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SUMM</td>
<td>Simple Urban Energy Balance Model for Mesoscale Simulation</td>
<td>Kanda et al. (2005a,b; 2007), Kawai et al. 2007, 2009)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TEB</td>
<td>Town Energy Balance</td>
<td>Masson (2000), Masson et al. (2002), Lemonsu et al. (2004)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TEB07</td>
<td>Town Energy Balance 7</td>
<td>Hamdi and Masson (2008)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TUF2D</td>
<td>Temperatures of Urban Facets 2D</td>
<td>Krayenhoff and Voogt (2007)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TUF3D</td>
<td>Temperatures of Urban Facets 3D</td>
<td>Krayenhoff and Voogt (2007)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>VUCM</td>
<td>Vegetated Urban Canopy Model</td>
<td>Lee and Park (2008)</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table 1:** Urban energy balance models with the number of versions and number of groups utilising each model participating in the comparison
### (a) Class details

<table>
<thead>
<tr>
<th>Fluxes included (F)</th>
<th>Cap</th>
<th>VL92</th>
</tr>
</thead>
<tbody>
<tr>
<td>All fluxes (a)</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>No Q_\text{H} (e)</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>No Q_\text{E} (f)</td>
<td>13</td>
<td>16</td>
</tr>
<tr>
<td>Neither Q_\text{H} nor Q_\text{E} (g)</td>
<td>0</td>
<td>5</td>
</tr>
</tbody>
</table>

### (b) Turbulent Flux Methods

<table>
<thead>
<tr>
<th>Resistance/Conductance (G)</th>
<th>B</th>
<th>V</th>
<th>B</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single layer (3)</td>
<td>19</td>
<td>4</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>Multilayer (4)</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Surface temp/moisture (Z)</th>
<th>B</th>
<th>V</th>
<th>B</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single layer air (3)</td>
<td>24</td>
<td>14</td>
<td>17</td>
<td>14</td>
</tr>
<tr>
<td>Multi layer air (4)</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Forcing height (F)</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

### (c) Urban morphology (L) and vegetation (V) combinations (see b)

<table>
<thead>
<tr>
<th>L</th>
<th>V</th>
<th>Q_\text{H}</th>
<th>Q_\text{E}</th>
<th>B</th>
<th>V</th>
<th>B</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td></td>
<td>113 (4)</td>
<td>113 (2)</td>
<td>9</td>
<td>11</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N (2)</td>
<td>N (2)</td>
<td>31</td>
<td>31</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11F (1)</td>
<td>11F (1)</td>
<td>33</td>
<td>33</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31F (2)</td>
<td>N (2)</td>
<td>33</td>
<td>33</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>33F (5)</td>
<td>N (5)</td>
<td>33</td>
<td>33</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>33S (3)</td>
<td>N (3)</td>
<td>33</td>
<td>33</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>33N (1)</td>
<td>N (1)</td>
<td>33</td>
<td>33</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td>L2</td>
<td></td>
<td>113 (2)</td>
<td>113 (2)</td>
<td>31</td>
<td>31</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td></td>
<td>N (2)</td>
<td>N (2)</td>
<td>33</td>
<td>33</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11F (1)</td>
<td>11F (1)</td>
<td>33</td>
<td>33</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>31F (2)</td>
<td>N (2)</td>
<td>33</td>
<td>33</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>33F (5)</td>
<td>N (5)</td>
<td>33</td>
<td>33</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>33S (3)</td>
<td>N (3)</td>
<td>33</td>
<td>33</td>
<td>33</td>
<td>33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>33N (1)</td>
<td>N (1)</td>
<td>33</td>
<td>33</td>
<td>33</td>
<td>33</td>
</tr>
</tbody>
</table>

### (d) Energy Balance Closure

<table>
<thead>
<tr>
<th>Closure forced by</th>
<th>Yes</th>
<th>No</th>
<th>VL92 actual closure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surface Temperature</td>
<td>6</td>
<td></td>
<td>9</td>
</tr>
<tr>
<td>AQ_\text{H} residual</td>
<td>17</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>31</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Characteristics used to classify the models (see Fig. 1). (a) Class code and number of models that have this capability (cap) and were applied this way for VL92. Classes with few models (*) are amalgamated for analysis. (b) Approaches used to simulate the built (B) and vegetated (V) turbulent heat fluxes (Q_\text{H}, Q_\text{E}). The numbers are for the VL92 runs. (c) Combined features (from b) GZA or ‘Other’ used in the turbulent flux modelling for the VL92 runs with the numbers in each class shown. (d) Energy balance closure approach. †1 is also 343. § 1 is 3P3 and 1 is 343.
Figure 1: Characteristics used to classify models (see Table 2).
2. The Characteristics of Urban Energy Balance Models

Urban energy balance models can be classified in a number of ways (see also Grimmond et al. 2009a), for example, they vary in terms of the fluxes they calculate (‘F’ in Fig. 1, Table 2). While all the models examined here calculate $K↑$, $L↑$, $Q^*$ and $Q_H$, some do not model either $Q_E$ or the $Q_F$, and some model neither. Here, a series of features are used to classify the approaches taken. Table 2 and Fig. 1 illustrate these and the former provides the numbers of models in each category. The illustrations also give each model class a reference in order to identify the category and its classification.

2.1 Vegetation and latent heat flux

(‘V’ in Fig. 1, Table 2)

A key decision in modelling an urban surface is whether or not vegetation (V) is simulated. A threefold classification is used here, where vegetation is:

- $V_n$: not considered;
- $V_s$: modelled using a “tile” scheme to represent the surface heterogeneity (e.g. Essery et al. 2003) which does not interact with other surface types until the first atmospheric level of a meso-scale model (e.g. Best et al., 2006); and
- $V_i$: ‘integrated’ into the modelled urban surface

The implication of not including vegetation is that there can be no latent heat except for periods immediately following rainfall. Some, even after rainfall, calculate no $Q_E$ whereas some account for dewfall and its later evaporation (Table 2b). For central business districts in many cities it may be reasonable to assume a negligible amount of vegetation and, hence, an absence of $Q_E$ associated with vegetated surfaces. However, in residential areas (e.g. suburban North America) extensive fractions of the surface are vegetated so the assumption of no urban $Q_E$ is unrealistic. Moreover, in many locations, extensive street cleaning can result in water being available for evaporation despite the lack of vegetation (e.g. Mexico City, Oke et al. 1999; Marseille, Grimmond et al. 2004).

The two classes of model that do incorporate vegetation differ in terms of the interactions which occur between the ‘built’ and ‘vegetated’ fractions (Table 2a, b, c). In the first case, ‘tiles’, ($V_s$, Fig. 1), models typically take advantage of traditional land surface schemes that have a wide variety of vegetation categories (e.g. Noilhan and Mahfouf 1996, Chen and Dudhia 2001, Essery and Clark 2003). Many have been extensively evaluated in the ‘Project for Intercomparison of Land-surface Parameterization Schemes’ (PILPS) (Henderson-Sellers et al. 1993, 2003, Irranejad et al. 2003) and other studies. Urban vegetation typically is more diverse than an individual vegetation class so a number of classes may be required (e.g. needleleaf and evergreen broadleaf trees) to ensure adequate representation. In the ‘tile’ approach, the ‘built’ and ‘vegetated’ fluxes are typically weighted by their respective plan area fractions to contribute to total fluxes (e.g. Lemonsu et al. 2004).

The ‘integrated’ case ($V_i$) is the most physically realistic as it allows for direct interaction of ‘built’ and ‘vegetated’ surfaces. This additional complexity may require increased computing resources and parameter values.

2.2 Anthropogenic heat fluxes

(‘AN’ in Fig. 1, Table 2)

The magnitude of $Q_F$ varies across a city. Typically it will be greatest in the densest part of the city (Oke 1988, Grimmond 1992, Ichinose et al. 1999). But even low absolute $Q_F$ values may be important where they exceed the radiative forcing (e.g. cloudy, cold winters with low solar input).

Similar to $Q_E$, not all models consider $Q_F$. The four general approaches are:
Ann: $Q_F$ is assumed to be zero, negligible or ignored;  
ANp: $Q_F$ is assumed to be a fixed amount that is required as specified input to the model, or is directly coded into the programme; 
ANi: $Q_F$ is calculated based on assumed internal building temperature; and 
ANm: $Q_F$ is calculated and incorporates internal heat sources from buildings, and/or mobile sources associated with traffic, and/or metabolism.

Models that calculate $Q_F$ typically include the heat related to internal heating of buildings as a minimum. A fixed temperature is assigned internally and this may, or may not, be allowed to vary seasonally or diurnally. Alternatively a fixed minimum (maximum) temperature is used so the internal temperature of the building may vary but within limits. The heat flux from traffic typically is based on assumptions about traffic flow, from vehicle counts. The models that calculate $Q_F$ in more detail, using a building energy model, mostly use the method of Kikegawa et al. (2003).

Beyond the internal temperature, the introduction of $Q_F$ requires consideration of where the heat is released or added to the atmosphere; for example, whether heat is added within or above the canyon.

2.3 Anthropogenic heat fluxes: temporal variation  
(‘T’ in Fig. 1, Table 2)  
$Q_F$ varies both diurnally and seasonally (e.g. Sailor and Lu 2004, Offerle et al. 2005, Pigeon et al. 2007, Lee et al. 2009), although only some models consider this. Models that prescribe a fixed value ($T_f$) are likely to provide too much $Q_F$ at night and insufficient quantities in the day; they will also not capture peak values normally associated with commuting (seasonally this peak may be associated with low solar radiation forcing). The inclusion of a diurnal and/or seasonal cycle ($T_v$) is more significant for certain applications when the modelled fluxes must be correct for specific short time periods. It is less significant when applications are not concerned with diurnal patterns.

2.4 Urban morphology  
(‘L’ in Fig.1, Table 2)  
Urban morphology affects radiative and turbulent heat exchanges. A number of approaches are used to capture these features, including:  
L1: Slab or bulk surface;  
L2: Single-layer approaches, which separate the surface into a ‘roof’ and ‘canyon’ (wall plus road); or  
L3: where the three facets (‘roof’, ‘wall’ and ‘road’) are treated separately; and  
L4-L7: Multiple-layer approaches, which divide one or more of the facets into layers or patches.  
Slab models represent the urban form as a flat horizontal surface with appropriate ‘bulk’ radiative, aerodynamic and thermal characteristics. This has the advantage of simplicity and reduced computational time and parameter requirements. Single layer models simplify the urban form to an urban canyon with a ‘roof’, ‘wall’ and a ‘road’. This allows for more realistic representations of radiative trapping and turbulent exchange (Masson 2000, Kusaka et al. 2001, Harman et al. 2004a, b, Lee and Park 2008). Parameter values are assigned for each facet and one set of energy exchanges per facets is modelled. Multi-layer schemes divide the walls into a number of vertical and/or the roof and road into a number of horizontal patches; each with their own parameter values and energy exchanges modelled. For some models this allows for variable building height, and for others even differing ‘roof’, ‘wall’ and ‘road’ characteristics. Note that the range of multilayer models L4-L7 is not exhaustive, rather it covers the range compared here.
2.5 Urban morphology: facets and orientation

(‘FO’ in Fig. 1, Table 2)

Models can be further sub-divided by how urban canyon morphology, specifically the number of facets and orientations, are dealt with. Models include: those which assume no facets (or orientation) and hence, a bulk (or slab) surface (FO1), those that assume one infinitely long canyon (FOn) and those which have infinitely long canyons that run in two cardinal directions (FOo). The canyons may be fixed in orientation and neglect shading or assume a random distribution of street canyons within the domain. Alternatively, the canyon may be modelled assuming two walls which have sunlit and shaded fractions that vary through the day and year. More realistic models also include an intersection between canyons (FOi), significantly increasing the number of the interactions with other facets that need to be computed.

2.6 Radiative fluxes: reflections

(‘R’ in Fig. 1, Table 2)

As $K_{\downarrow}$ and $L_{\downarrow}$ are provided, it is the $K_{\uparrow}$ and $L_{\uparrow}$ that are modelled. Beyond the morphology, and therefore the degree of detail needed for the surface parameters, the major differences relate to the number of reflections assumed: R1: Single; Rm: Multiple; and Ri: Infinite.

The single reflection model is the least computationally intensive and used in both slab and single layer models. Models which simulate multiple reflections include both single layer and multiple layer models. Infinite reflections may be accounted for by slab, single layer and multi-layer models.

For long-wave radiation, slab models determine one surface temperature, whereas for facet-specific models, multiple surface temperatures are calculated (Table 2b,c). The surface temperatures then provide the forcing for $Q_H$ and $\Delta Q_S$.

2.7 Radiative fluxes: albedo and emissivity

(‘AE’ in Fig.1, Table 2)

The albedo and emissivity values that determine the radiative fluxes may either be defined as a single value (bulk, AE1), as two facets, similarly to the L2 category (AE2),or may consist of combinations of various facets, analogous to the L3 (or greater) category of models (AEf).

2.8 Storage heat flux

($\Delta Q_S$) (‘S’ in Fig. 1, Table 2)

The $\Delta Q_S$ is significant in urban areas given the materials and morphology of the urban surface (Grimmond and Oke 1999a). In urban models, it is determined in the following ways:

Sr: difference or residual of the energy balance;
Sc: solution of the heat conduction equation by dividing the facets into a number of thickness layers; and
Sn: function of $Q^*$ and surface characteristics.

All three methods are used by slab or bulk models (Table 3). For all three the ability to model the outgoing long-wave radiation will impact the values obtained given the common need to model surface temperature.

For those models in which heat storage is calculated as the residual of the surface energy balance, assumptions as to which fluxes are included (specifically $Q_F$, $Q_E$) are important. The second method, the solution of the heat conduction equation, is used extensively by slab, single and multiple layer models. It requires various parameters for each (sub-) facet, including: number of layers, layer thickness, thermal conductivity and volumetric heat capacity of the various layer materials (Table 4). The number of layers resolved varies between 1 and 48, and may be of fixed or variable thickness. Currently, none account for changing water content of built materials associated with rain, so the material parameters are static. Some solve the heat conduction equation using the
force-restore method, while others solve the one-dimensional heat conduction equation.

The third approach is to use a fraction of \( Q^* \) (Sn). Some models take into account the diurnal pattern of the flux through the objective hysteresis model (Grimmond et al. 1991).

### 2.9 Other features

The following characteristics are not explicitly used to classify the models in this evaluation but are presented here because of differences between models. They do not necessarily result in the models being grouped differently to the classifications above; that is, models fall into some common groupings across model classes (Table 2, 3).

### 2.10 Turbulent sensible and latent heat fluxes

**Turbulent sensible (\( Q_H \)) and latent heat (\( Q_E \)) flux**

Typically surface resistance (or its inverse, conductance) schemes are used to model \( Q_H \) and \( Q_E \) (‘G’ in Fig. 1, Table 2b). Depending on urban morphology, these consist of either single (G3) or multiple (G4) resistance networks which account for the number of facets and layers that are resolved. Bulk models (G1) have the simplest resistance network (Table 2b, Fig. 1). A wide range of resistance schemes are used (e.g. Rowley et al. 1930, Clarke 1985, Zilitinkevich 1995, Guilloteau, 1998, Harman et al. 2004b). To determine the resistance the wind profile within/above the canyon, roughness length and displacement length or drag coefficients and atmospheric stability may be taken into account. Drag is either not considered or is calculated using roughness length, exponential wind profile, or distributed drag. Exchange between the canopy air and building surfaces may be parameterized by a roughness length approach or distributed sources of heat (generally in conjunction with a distributed drag approach).

The number of temperatures resolved, which drive the gradients, varies both for the surface and the air (within the canyon) (‘Z’ and ‘A’ respectively in Table 2b, Fig.1), and these are related to morphology and the number of vertical layers in the model. Many assume Monin-Obukhov similarity holds which may not be applicable within the urban canyon or within the roughness sublayer (Roth 2000). However, given the lack of well-tested alternatives, currently this may be the most appropriate approach.

As \( Q_H \) is calculated typically using surface temperature to force the gradient, a balance is inherent in the solution of surface temperature between the \( L \uparrow \), \( Q_H \) and heat conduction. Depending on the model objective, performance may be improved for one flux at the expense of another. Models that use a combination method (P, Penman-Monteith or combination-type approach) do not need to determine the surface temperature to calculate \( Q_H \), but still need to allow for the transport of heat away from the surface.

The approaches taken to model resistances (G), surface temperature (Z) and air temperature (A) result in a large number of combinations (Table 2c, expressed in GZA order). Here they are shown relative to the urban morphology classes (L1-L7, Fig. 1) and the vegetation class (VN, VS, VI). The approach taken for each turbulent flux \( (Q_H, Q_E) \) for the built (B) and vegetated (V) part of the surface are shown. It is clear that the earlier classifications (Table 2a) do not produce common characteristics for these fluxes. Given the wide range of approaches these are not investigated in further detail in this paper. Subsequent analysis of a larger data set will investigate this. For the calculation of \( Q_H \) for the built (B) fraction of the surface, the two most common classes of the nine different combinations are:

- 333: single layer resistance (G3), surface temperature (Z3) and air temperature (A3);
- 113: bulk resistance per facet (G1) and surface temperature (Z1) and a single air temperature (A3);

For the vegetated surfaces the two most common classes for \( Q_H \) are:

- N: \( Q_H \) is not calculated; and 113.
- For \( Q_E \) from built surfaces the predominant classes are N, 113 and 333; but also of note are those models that account for the evaporative loss of water in one time step immediately following
### Table 3: Analysis of the individual characteristics (Table 2) gives this combination of approaches: (top & right) are the capability of the models and (bottom & left) is how the models were applied for the VL92 runs analysed here. See Fig. 1 and Table 2 for explanation of class codes. For example, with respect to the capability of the 33 models: five which model vegetation in a separate tile (Vs) have slab morphology (L1), while in application to the VL92 dataset, only two were run this way.
<table>
<thead>
<tr>
<th>Parameters by Capabilities (models, VL92)</th>
<th># models (capability)</th>
<th># models (VL92)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Radiative</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Albedo (-)</td>
<td>24,22,19,9,6,6,4</td>
<td>24,22,19,9,6,6,4</td>
</tr>
<tr>
<td>VIS, UV and NIR albedo of vegetation (-)</td>
<td>1,03</td>
<td>2,14</td>
</tr>
<tr>
<td>Emissivity (-)</td>
<td>23,21,19,5,6,5,2</td>
<td>23,21,19,5,6,5,2</td>
</tr>
</tbody>
</table>

| Roughness                                |                       |                 |
| Roughness length above canyon (m)        | 10,8                  | 10,10           |
| Material roughness length for heat (m)   | 14,16,6               | 4,16,6          |
| Material roughness length for momentum (m) | 16,14,6             | 17,14,6         |
| Effective roughness length for heat (m)  | 0                     | 5,0,2           |
| Effective roughness length for momentum (m) | 4,5                  | 5,2             |
| Roughness length of grass/tree surfaces (m) | 1                    | 2               |
| Bulk roughness length (m)                | 23,21,19,5,6,5,2      | 23,21,19,5,6,5,2|
| Sub-layer Stanton number (+)             | 2,1,1                 | 2,1,1           |

| Temperature §                            |                       |                 |
| Mean internal building temperature (K)   | T_w                       | 6               |
| Deep temperature (K)                      | 6,6                       | 6,8             |
| Facet temperature (K)                     |                           | 6,6,5,4         |
| Min; max internal building temp (K)       | T_min; max internal       | 6,3             |
| temperature                               |                          | 6,3             |

| Vegetation / soil specific                |                       |                 |
| Vegetation wilting point (m²⁻¹)          | S_wh                    | 9               |
| Rooting depth of grass/tree (m)          | d_r                       | 3,3             |
| Minimum stomatal resistance (m⁻¹)       | R_s                      | 5               |
| Leaf area index of vegetation within the urban canyon (-) | LAI                     | 6               |
| Vegetation thermal inertia (J m⁻¹ K⁻¹ s⁻¹) | T_v                      | 1               |
| Parameters used in radiation stress function (-) | R_s                    | 3               |
| Parameter used in vapour pressure deficit function (-) | P_s                     | 3               |
| B parameter (+)                          | B                        | 12              |
| Sand/clay/loam/quartz content of soil (+) | S_s                      | 6,4,3,1         |
| Maximum vegetation canopy water holding capacity (mm) | M_s                      | 6,3             |
| Optimum temperature in temperature stress function (K) | T_opt                    | 1               |
| Coef. for maximum interception water storage capacity (-) | L_s                      | 1               |
| Ecosystem respiration parameter (-)      | E_s                      | 1               |
| Ratio d(biomass)/d(LAI) (+)              | d/d                      | 1,2             |
| e-folding time for senescence (-)        | E_f                      | 1               |
| Cuticular conductance (-)                | C_s                      | 1               |
| Maximum air saturation deficit (100 g kg⁻¹) | A_s                      | 1               |
| Leaf area ratio sensitivity to nitrogen (m² kg⁻¹) | S_s                     | 2               |
| Lethal minimum value of LA ratio (m² kg⁻¹) | L_s                      | 1               |
| Nitrogen concentration of biomass (m² kg⁻¹) | N_s                      | 2               |
| Root fraction (%)                        | R_f                      | 2               |
| Tree coverage (%)                        | T_c                      | 1               |
| Sunny spots (%)                          | S_s                      | 1               |
| Canopy solar absorptivity (-)            | C_s                      | 1               |
| Canopy solar transmissivity (-)          | C_t                      | 1               |
| Canopy Thermal time constant (-)         | E_c                      | 1               |
| Tree evaporative cooling coefficient (-) | E_f                      | 1               |

| Moisture availability                    |                       |                 |
| Moisture availability (m³ m⁻³)            | M_w                      | 3,7             |
| Critical normalized soil water content for stress (m s⁻¹) | S_w                      | 13              |
| Air dry soil moisture content limit (m s⁻¹) | S_m                     | 2               |
| Soil suction experienced in the soil at saturation (m) | S_s                     | 3               |
| Maximum soil moisture content (field capacity) (m² m⁻³) | M_max                    | 9               |

| Table 4: Urban parameters used by the number of models indicated for the VL92 runs and also those indicated as being applicable for the models (i.e. capability, cap) (# models: number of models which use the parameters in column 3, respectively). Each subscript refers to a separate parameter: f = roof; r = road; w = wall; v = pervious; t = tree; H = building; g = grass, |
precipitation with a fixed rate of evaporation (E). For $Q_e$ from vegetated surfaces the predominant classes are also N and 113. Two models which do simulate $Q_H$ and $Q_e$ for vegetated area, account for evaporation from soil moisture only and not the loss of water through vegetation. In these cases the soil temperature and moisture profile are calculated using the approach of Tremback and Kessler (1985). In urban areas bare soil is rare with some sort of vegetation most likely to be present.

2.11 Energy balance closure

Not all models explicitly force or check that they have energy balance closure (i.e. that equation (1) holds, Table 2d). Lack of closure may result from numerical instabilities or lack of precision in the code, from a lack of evaluation, or from inconsistent assumptions. Closure may be forced in a number of ways: through the calculation of $\Delta Q_S$ at the end of each time step as a residual, by updating the surface temperature of the facets, or by restricting the turbulent heat fluxes to the available energy ($Q^* - \Delta Q_S$). Closure is an important issue when the land surface scheme is part of a long term climate model simulation; without it, there may be long term bias in the model.

2.12 Anthropogenic water flux and other capabilities

Water can be added to the urban environment by human activity. Water is released by combustion processes, cooling towers and by people; equivalent to the $Q_F$ release (anthropogenic latent heat flux). One model takes into account the loss of water through perspiration ($Q_e$). Given there are very few estimates of this term (Heiple and Sailor 2008, Moriwaki et al. 2008), and it is likely to be small in many settings, it is not surprising that it has not been included in most models. The term may be important in very dry areas (e.g. high latitude cities in winter, hot dry cities) and in areas with excessive air conditioning. The second significant source of water comes via the pipe network, most typically as irrigation (e.g. garden sprinkling) or broken water pipes. In many suburban areas, if gardening is a common residential activity, this can be a large additional source of water relative to precipitation, especially during the summertime (e.g. Grimmond and Oke 1986, Grimmond et al. 1996). Estimating this component requires assumptions in the algorithms and/or the input data to define: (1) how much, and when, water is applied to the area and (2) where in the area it is released (e.g. to all vegetated surfaces or just to irrigated grass). The representation of this source is important (e.g. Mitchell et al. 2001) but has been considered in few models (e.g. Grimmond and Oke 1991).

The presence of snow-cover will influence the energy balance of urban regions, affecting the albedo and, during periods of snow melt, acting as a significant sink for latent energy (Lemonsu et al. 2008). For models with facets, the energy budgets of horizontal surfaces (roof and road) will be the most significantly affected, with additional energy budgets for these surfaces being necessary (Masson, 2000).

2.13 Model uniqueness

Using the 31 individual characteristics to classify the models compared (Table 2a), 26 unique combinations occur (Table 3). This varies between model capability and actual use (demonstrated here for a data set termed ‘VL92’, see section 4). For example, 21 models have the capability to account for $Q_F$ but only seven utilize this capability for the VL92 application. Although there are preferred approaches (e.g., $Q_F$ $T_v$ over $T_f$), there is a notable diversity; models that have a similar approach for one aspect frequently use quite different approaches for other model components.
3. Model Inputs

Inputs of three general types are required to model urban areas: (1) site parameters to describe the surface morphology and materials; (2) time series of atmospheric or forcing variables as boundary conditions; and (3) initial thermodynamic and moisture state conditions. The complexity of urban areas and diversity of surface description methods in the 33 models results in more than 145 (or > 200 if individual layer values are considered) different parameters and state variables being needed for all of the models. The parameters fall into nine broad classes (Table 4). Some parameters, which are unique to individual models, can be derived from more basic parameters (Table 5, Fig. 2). Given the large effort needed to collect these data over the wide range of urban areas globally, or even within individual countries (e.g. Feddema et al. 2005, Ching et al. 2009), we encourage model developers to use common parameters. Also, it is important that the parameters are clearly defined and not open to misinterpretation (Loridan et al. 2009).

Morphometric parameters vary greatly, using either basic information (e.g. height, width) from which required parameters are calculated (e.g. canyon aspect ratio, sky view factor), or the ‘higher’ level parameters as the inputs. Table 4 lists basic parameters from which higher level parameters can be calculated.

Urban material related parameters are required to account for radiative transfer (e.g. albedo, emissivity) and thermal characteristics. Because of the different ways to describe the surface (Table 2, Fig. 1), there are varying numbers of models that use particular parameters (Table 4). All models use some form of albedo but this may be a single bulk albedo ($\alpha_b$), or albedos for the roof ($\alpha_f$), wall ($\alpha_w$) and road ($\alpha_e$) etc. Thermal properties are specified explicitly either relative to mass (specific heat capacity) or volume (volumetric heat capacity), or implicitly from model ‘look-up’ tables.

As noted in Table 2, $Q_F$ is dealt with in a variety of ways. For those using a fixed value, a model parameter has to be specified or alternatively, internal building temperature may need to be specified (in Table 4, it is included under Temperatures but could be specified under $Q_F$).

Temperatures are required for many models. These may be initial state conditions (e.g. facet temperatures) which will evolve during the run, or require model spin-up of sufficient time, or may be fixed for the duration of the run (e.g. deep soil temperature). In many applications, it is likely to be difficult to have realistic or observed values to meet the need for the temperature profile within a building or the soil to be prescribed. This may mean that some models require a long initialization period (spin-up) to ensure that the temperature profiles are stable and representative of expected conditions.

For the models that use a vegetation tile, all the parameters required are not summarised in Table 4. Parameter values, based on class selection, have been determined for extensive non-urban vegetated areas, and are assigned through model ‘look-up’ tables. Model users have selected the
vegetation class (e.g. grassland, deciduous or evergreen woodland and/or bare soil) that they think is most appropriate in relation to the urban region they are modelling.

Soil moisture characteristics require both initial values and fixed parameters. These state variables have similar constraints and implications to that of the temperature. As urban areas often have disturbed soils and additional materials mixed into the media, it may mean that adoption of rural soil physical properties for parameters is not appropriate.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Derivation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Fundamental morphometric parameters (m)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean building height</td>
<td>$z_H$</td>
<td></td>
</tr>
<tr>
<td>Mean building length</td>
<td>$L_{XY}$</td>
<td>$L_{XY} = \frac{\sum_{i=1}^{n} (L_i + L_{yi})}{2n}$</td>
</tr>
<tr>
<td>Mean building width</td>
<td>$L_{XY}$</td>
<td>$L_{XY} = \frac{\sum_{i=1}^{n} (L_i + L_{yi})}{2n}$</td>
</tr>
<tr>
<td>Mean canyon width</td>
<td>$W_{XY}$</td>
<td>$W_{XY} = \frac{\sum_{i=1}^{n} (W_i + W_{yi})}{2n}$</td>
</tr>
<tr>
<td>Mean block length</td>
<td>$D_{XY}$</td>
<td>$D_{XY} = L_{XY} + W_{XY}$</td>
</tr>
<tr>
<td>Mean building length</td>
<td>$L_{XY}$</td>
<td>$L_{XY} = \frac{\sum_{i=1}^{n} (L_i + L_{yi})}{2n}$</td>
</tr>
<tr>
<td>Mean building width</td>
<td>$L_{XY}$</td>
<td>$L_{XY} = \frac{\sum_{i=1}^{n} (L_i + L_{yi})}{2n}$</td>
</tr>
<tr>
<td>Mean canyon width</td>
<td>$W_{XY}$</td>
<td>$W_{XY} = \frac{\sum_{i=1}^{n} (W_i + W_{yi})}{2n}$</td>
</tr>
<tr>
<td>Mean block length</td>
<td>$D_{XY}$</td>
<td>$D_{XY} = L_{XY} + W_{XY}$</td>
</tr>
<tr>
<td>Mean building separation</td>
<td>$XYW$</td>
<td></td>
</tr>
<tr>
<td>Plan area ratio</td>
<td>$\lambda_p$</td>
<td>$\lambda_p = \frac{L_{XY}^2}{D_{XY}^2}$</td>
</tr>
<tr>
<td>Canyon aspect ratio</td>
<td>$\lambda_s$</td>
<td>$\lambda_s = \frac{z_H}{W_{XY}}$</td>
</tr>
<tr>
<td>Canyon height to width ratio</td>
<td>$\lambda_s$</td>
<td>$\lambda_s = \frac{z_H}{W_{XY}}$</td>
</tr>
<tr>
<td>Roof area ratio</td>
<td>$\lambda_s$</td>
<td>$\lambda_s = \frac{z_H}{W_{XY}}$</td>
</tr>
<tr>
<td>Building coverage ratio</td>
<td>$\lambda_s$</td>
<td>$\lambda_s = \frac{z_H}{W_{XY}}$</td>
</tr>
<tr>
<td>Road area fraction</td>
<td>$\lambda_s$</td>
<td>$\lambda_s = \frac{z_H}{W_{XY}}$</td>
</tr>
<tr>
<td>Building frontal density</td>
<td>$\lambda_s$</td>
<td>$\lambda_s = \frac{z_H}{W_{XY}}$</td>
</tr>
<tr>
<td>Frontal area index</td>
<td>$\lambda_s$</td>
<td>$\lambda_s = \frac{z_H}{W_{XY}}$</td>
</tr>
<tr>
<td>Wall to non-built horizontal area</td>
<td>$\lambda_{wnb}$</td>
<td>$\lambda_{wnb} = \frac{2(L_z + L_y)z_H}{W_y D_z + W_z L_y}$</td>
</tr>
<tr>
<td>(b) Derived morphological parameters</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Density (kg m$^{-3}$)</td>
<td>$\rho$</td>
<td>Thermal conductivity (W m$^{-1}$ K$^{-1}$)</td>
</tr>
<tr>
<td>Specific heat (J kg$^{-1}$ K$^{-1}$)</td>
<td>$c$</td>
<td>Volumetric heat capacity (J m$^{-3}$ K$^{-1}$)</td>
</tr>
</tbody>
</table>

Table 5: (a) Fundamental morphometric parameters (units: m) that can be used to derive (b) dimensionless morphometric parameters. (c) Thermal parameters. Note many different names are used for the same parameters. Refer to Fig. 2 for further definitions.
4. The International Urban Energy Balance Model Project

The methodology adopted here follows that of PILPS (Henderson-Sellers et al. 1993) which provided insight into both the models and real world processes. This allows the relative importance of key parameters to be determined and an assessment of the level of complexity required to produce reliable results. The International Urban Surface Energy Balance Model Comparison Project has been endorsed by the GEWEX Global Land-Atmosphere System Study (GLASS) and World Meteorological Organization Expert Team on Urban and Building Climatology.

The procedure for the comparison requires individual modelling groups (users and/or developers) to run their model(s) ‘offline’ using forcing data provided for the ‘top’ of the model, as would be provided by an atmospheric model (Fig. 3). This implies that parameter values should be representative of the observational footprint (see discussion in Masson et al. 2002). There is no feedback to larger scale conditions within the modelling domain, so no larger scale advection can occur, as would be present in a meso-scale or larger scale model. The temporal resolution of analysis is typically 30 or 60 min, but individual models may be run at higher temporal resolution (1.5 – 300 s) and then average or sample their data back to the specified time interval of analysis (60 minutes). The spatial scale for both the measurements and models is the local or neighbourhood scale (10^2-10^4 m). However there is no actual grid size because the models are run in single column mode. The observed fluxes and the forcing data are taken from tall towers which have the sensors located above the roughness sublayer (Grimmond and Oke 1999b, Roth 2000, Masson et al. 2002, Grimmond et al. 2004, Grimmond 2006). This height is equivalent to being above or at the blending height and is typically taken as the first atmospheric layer in meso-scale or larger scale models (Fig. 3).

![Figure 3](image)

**Figure 3:** Urban land surface schemes simulate exchanges between the urban surface and the first layer of larger scale atmospheric models. The observed fluxes, and the forcing data, are representative of the same level since they are above the roughness sub layer or blending height.

The rationale for offline simulation is that although larger-scale circulation models may be accurate at the macro-scale, their outputs will often be incompatible with those required as inputs to meso-scale urban surface models (Pitman et al. 1990). Equally, running such models offline prevents feedbacks between climate and land surface, meaning that the sensitivity of the land surface schemes themselves can be examined while the overlying atmospheric conditions are effectively held fixed (Wilson et al. 1987, Henderson-Sellers and Dickinson 1992).
Table 6: Summary of the mean, maximum and minimum statistical performance (see Table 7 for definitions of statistics) across 33 models when compared to the VL92 data set for all hours ($n=312$ hours). Also, RMSE statistics are displayed for the first run of the models. RMSE values from previous evaluations using VL92 data for all hours (note these are not directly comparable as different time periods and forcing data are used in some cases): $^1$Grimmond and Oke (2002), $^2$Masson et al. (2002) for periods 223-236, 225-231, 232-236 respectively * combined $Q_H + Q_E$, $^3$Best et al. (2006) tile 1 and 2 respectively, $^4$Krayenhoff and Voogt (2007) LI1 (original simulation), and LI2 (with parameter adjustments) respectively for TUF3D for 0300-2100 on day 227, $^5$Oleson et al. (2008b) for periods 225-231 & 232-236 respectively.

<table>
<thead>
<tr>
<th>Statistic</th>
<th>$Q^*$</th>
<th>$Q_E$</th>
<th>$\Delta Q$</th>
<th>$Q_S$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_{obs}$ (W m$^{-2}$)</td>
<td>Max 193.6</td>
<td>208.5</td>
<td>83.9</td>
<td>30.7</td>
</tr>
<tr>
<td></td>
<td>Min 84.4</td>
<td>49.8</td>
<td>-15.3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Mean 133.7</td>
<td>113.6</td>
<td>13.3</td>
<td>7.4</td>
</tr>
<tr>
<td>$\sigma_{obs}$ (W m$^{-2}$)</td>
<td>Max 268.6</td>
<td>197.4</td>
<td>187.3</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Min 153.8</td>
<td>67.7</td>
<td>41.5</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Mean 231</td>
<td>120.5</td>
<td>119.2</td>
<td>8.8</td>
</tr>
<tr>
<td>$r^2$</td>
<td>Max 0.99</td>
<td>0.85</td>
<td>0.88</td>
<td>0.39</td>
</tr>
<tr>
<td></td>
<td>Min 0.85</td>
<td>0.61</td>
<td>0.45</td>
<td>0.01</td>
</tr>
<tr>
<td></td>
<td>Mean 0.98</td>
<td>0.8</td>
<td>0.79</td>
<td>0.25</td>
</tr>
<tr>
<td>RMSE (prior runs) (W m$^{-2}$)</td>
<td>1 - 49</td>
<td>20</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2 59.57,59</td>
<td>76.97, 50*</td>
<td>91, 105, 66</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>3 69.71</td>
<td>56,43</td>
<td>103, 81</td>
<td>27.24</td>
</tr>
<tr>
<td></td>
<td>4 40.2,31.1</td>
<td>138.5, 107.4</td>
<td>109.9, 98.1</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>5 34,34</td>
<td>81,49</td>
<td>88,59</td>
<td>16,23</td>
</tr>
<tr>
<td>$^*RMSE$ first run (W m$^{-2}$)</td>
<td>Max 177.9</td>
<td>233.3</td>
<td>311.4</td>
<td>157.4</td>
</tr>
<tr>
<td></td>
<td>Min 28.4</td>
<td>39.3</td>
<td>49.1</td>
<td>17.2</td>
</tr>
<tr>
<td></td>
<td>Mean 58.4</td>
<td>95.5</td>
<td>87.8</td>
<td>30.3</td>
</tr>
<tr>
<td>RMSE final run (W m$^{-2}$)</td>
<td>Max 92.3</td>
<td>183.1</td>
<td>115.7</td>
<td>31.5</td>
</tr>
<tr>
<td></td>
<td>Min 22.1</td>
<td>36.8</td>
<td>49.1</td>
<td>17.2</td>
</tr>
<tr>
<td></td>
<td>Mean 47</td>
<td>81.7</td>
<td>77.8</td>
<td>23</td>
</tr>
<tr>
<td>RMSE$_{\delta}$ (W m$^{-2}$)</td>
<td>Max 81.9</td>
<td>163.8</td>
<td>111.5</td>
<td>26.1</td>
</tr>
<tr>
<td></td>
<td>Min 4.2</td>
<td>6.8</td>
<td>16</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>Mean 30.3</td>
<td>58.9</td>
<td>54.8</td>
<td>19.8</td>
</tr>
<tr>
<td>RMSE$_{\beta}$ (W m$^{-2}$)</td>
<td>Max 80.7</td>
<td>81.8</td>
<td>79.3</td>
<td>27.6</td>
</tr>
<tr>
<td></td>
<td>Min 18.1</td>
<td>32.2</td>
<td>15.7</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Mean 33.6</td>
<td>53</td>
<td>50.6</td>
<td>7.4</td>
</tr>
<tr>
<td>MAE (W m$^{-2}$)</td>
<td>Max 76.6</td>
<td>136.7</td>
<td>89.7</td>
<td>21.4</td>
</tr>
<tr>
<td></td>
<td>Min 15.4</td>
<td>24.7</td>
<td>33.1</td>
<td>11.5</td>
</tr>
<tr>
<td></td>
<td>Mean 37</td>
<td>57.1</td>
<td>57.5</td>
<td>15.6</td>
</tr>
<tr>
<td>MBE (W m$^{-2}$)</td>
<td>Max 62.4</td>
<td>136.7</td>
<td>41.4</td>
<td>15.2</td>
</tr>
<tr>
<td></td>
<td>Min -46.9</td>
<td>-22</td>
<td>-57.7</td>
<td>-15.6</td>
</tr>
<tr>
<td></td>
<td>Mean 2.7</td>
<td>44.1</td>
<td>-30.9</td>
<td>-8.3</td>
</tr>
<tr>
<td>$d$</td>
<td>Max 1</td>
<td>0.95</td>
<td>0.96</td>
<td>0.78</td>
</tr>
<tr>
<td></td>
<td>Min 0.94</td>
<td>0.66</td>
<td>0.58</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>Mean 0.99</td>
<td>0.86</td>
<td>0.88</td>
<td>0.54</td>
</tr>
</tbody>
</table>
The forcing data provided to participants with were collected from a light industrial site in Vancouver, BC, Canada (termed here ‘VL92’) (Voogt and Grimmond 2000, Grimmond and Oke, 2002). All observational data have measurement errors. These are associated with instrumental errors, instrument siting, fetch, flux corrections, lack of energy balance closure, and neglected terms etc. (e.g. Offerle et al. 2005, Grimmond 2006). This dataset was chosen as it has been used previously by a number of groups to evaluate their models (Grimmond and Oke 2002, Masson et al. 2002, Best et al. 2006, Krayenhoff and Voogt 2007, Oleson et al. 2008b). This meant that parameter values were reasonably well known. Also, the observed fluxes were provided so no model/group had an advantage from previous knowledge of this data.

The observations used in the evaluation consist of $Q^*$, $Q_H$ and $Q_E$ plus $\Delta Q_S$ determined as a residual (Grimmond and Oke 1999a). During the observations (14 days in August 1992) the area was in drought and there was an irrigation ban in the city that was adhered to (Grimmond and Oke 1999c). The area is characterised by little vegetation (< 5% plan area cover) and the soil moisture was very low at the time of data collection (Grimmond and Oke 1999a,c), making $Q_E$ at this site small compared to the other fluxes (Table 6). The summertime conditions are expected to be associated with low $Q_F$ as the area did not have extensive use of air conditioning or other significant sources of $Q_F$. This would be expected to be more significant in the winter but is not considered here as no observational data were available. The purpose of this comparison is not to identify the

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Description/equation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{x}$</td>
<td>Mean $\bar{P} = \frac{\sum_{i=1}^{n} P_i}{n}$</td>
</tr>
<tr>
<td>$\sigma$</td>
<td>Standard deviation $\sigma_P = \sqrt{\frac{\sum_{i=1}^{n} (P_i - \bar{P})^2}{n-1}}$</td>
</tr>
<tr>
<td>$R$</td>
<td>Correlation coefficient (Pearson’s) $R = \frac{\sum_{i=1}^{n} O_i P_i - \frac{\sum_{i=1}^{n} O_i \sum_{i=1}^{n} P_i}{n}}{\sqrt{\left(\sum_{i=1}^{n} O_i^2 - \left(\frac{\sum_{i=1}^{n} O_i}{n}\right)^2\right)\left(\sum_{i=1}^{n} P_i^2 - \left(\frac{\sum_{i=1}^{n} P_i}{n}\right)^2\right)}}$</td>
</tr>
<tr>
<td>$R^2$</td>
<td>Coefficient of Determination $R^2$</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root mean squared error $RMSE = \left[n^{-1} \sum_{i=1}^{n} (e_i)^2 \right]^{0.5}$</td>
</tr>
<tr>
<td>RMSE_s</td>
<td>Systematic RMSE $RMSE_s = \left[n^{-1} \sum_{i=1}^{n} (\hat{P}_i - O_i)^2 \right]^{0.5}$</td>
</tr>
<tr>
<td>RMSE_u</td>
<td>Unsystematic RMSE $RMSE_u = \left[n^{-1} \sum_{i=1}^{n} (P_i - \hat{P}_i)^2 \right]^{0.5}$</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean Absolute Error $MAE = n^{-1} \sum_{i=1}^{n}</td>
</tr>
<tr>
<td>MBE</td>
<td>Mean Bias Error $MBE = n^{-1} \sum_{i=1}^{n} e_i = \sum \bar{P} - \bar{O}$</td>
</tr>
<tr>
<td>$d$</td>
<td>Index of Agreement $d = 1 - \frac{\sum_{i=1}^{n} (e_i)^2}{\sum_{i=1}^{n} (</td>
</tr>
</tbody>
</table>

Table 7: Statistics used to analyse model performance (Willmott, 1981; Jacobson, 1999): $P_i$ and $O_i =$ predicted and observed values; $n =$ number of data points; $e_i = P_i - O_i$ $\hat{P} = a + bO_i$ (where $a$ and $b$ are the intercept and slope of regression line between $O$ and $P$).
best model, but to understand model errors related to the type of approach taken (Table 2, Fig. 1). Each model was assigned a random identifier number which is used in the subsequent analysis of the results to ensure anonymity. The returned simulation data from each of these models were used to perform a series of statistical analyses to evaluate model performance (Table 7).

5. Results from VL92

Using the VL92 dataset, 33 different models/versions of models were analysed (Table 1). Modelling groups assigned parameter values and initial state conditions they thought appropriate. Of the 33 participants, 20 chose to re-run their models subsequent to their initial submission and based on developments of their models during the period of the model comparison, thereby improving performance. Of those who did 16, 3, and 1 groups respectively re-ran their models once, twice and three times with a decrease in the root mean square error (RMSE) in all cases except for the minimum values for $Q_E$ and $\Delta Q_S$ which remain the same (Table 6). The remainder of this paper evaluates the performance based on the ‘final’ runs only.

As noted, this site has been used to evaluate model performance in previous studies (Table 6). These evaluations are not directly comparable to the current data as the same forcing data were not used in all the studies, and the time periods are not consistent, unlike the current comparison where all models have followed an identical protocol. However, comparing those results to the ‘final’ runs presented here we can see that the results are similar. As with the overall cohort of models participating in the International Urban Model Comparison, there is some suggestion that model performance may have improved in the current (‘final’) runs.

For $Q^*$ the models, on average, have a smaller systematic RMSE (RMSE$_S$) than unsystematic RMSE (RMSE$_U$) (Table 6). However, the maximum RMSE$_S$ (81.9 W m$^{-2}$) is the same order of magnitude as the maximum RMSE$_U$ (80.7 W m$^{-2}$) suggesting there are problems that could be fixed, for example by changing parameter values. For $Q_H$ the mean and maximum RMSE$_S$ are larger than the RMSE$_U$, also suggesting that model results might be improved.

The ranked performance of the individual models, based on RMSE calculated for the 312 h dataset, for the four fluxes, is shown in Fig. 4. No individual model performs ‘best’ or ‘poorest’ for all fluxes. For each flux, when models are ordered from ‘best’ to ‘poorer’ performance, in the better performing models there are small differences in RMSE. However, there is a point of step-drop in performance: for $Q^*$ five models performing less well; for $Q_H$ 15 models show distinctly poorer performance.

The encouraging performance for $Q_E$, with small RMSE values and only two models performing noticeably poorer, is a function of its small flux (Table 6). When a normalized Taylor (2001) plot is considered (Fig. 4e-h), where the ideal model would fall at the square (the observations), $Q_E$ is the least well modelled (Fig. 4h). For $Q^*$, the models cluster most closely to the observed value, except for the five outliers already identified. Again for $Q^*$, all models have a correlation coefficient ($r$) of greater than 0.95 except for one, which has an $r$ value of over 0.9.

Interestingly, there is less of a step drop in $\Delta Q_S$ model performance but an almost constant correlation coefficient for all models (~0.9). $Q_H$ also has an almost constant correlation coefficient for all models (~0.9). Based on the index of agreement ($d$), on average model performance is best for $Q^*$, followed by $\Delta Q_S$, $Q_H$ and $Q_E$ (Table 6). This ranking is retained when the best overall performance (maximum $d$) of any model for each flux is considered.

Models need to respond to changes in exchange processes through the course of the day. Of interest, for example, is whether they resolve peak radiant and turbulent heat fluxes during the day as well as fluxes at night when shortwave radiation does not need to be considered. When the data are analysed by time of day, RMSE is larger during the day (Fig. 5) as expected because of its larger absolute magnitude. Fig. 5 shows results for three time periods: (a) day (1 h after $Q^* \geq 0$ W m$^{-2}$), (b) night (1 h after $Q^* \leq 0$ W m$^{-2}$), and (c) transition (remaining hours when $Q^*$ is going
The five models with the largest RMSE for daytime $Q^*$ (Fig. 5), are the same as those for all hours (Fig. 4), although the ranked order differs slightly. The transition hours are particularly problematic for these models. The two poorest performing in the day time are among the six-poorest performing at night.

The observed fluxes of $Q_H$ may be underestimated on some occasions due to advection caused by sea breezes (Masson et al. 2002). For $Q_H$ the daytime errors are largest. At night the models generally do well almost across the board but the absolute values of the fluxes are smaller (Fig. 5). The daytime RMSE for $Q_H$ is larger than for $Q^*$ for all models. The RMSEs tend to be greater for $Q_H$ than for $Q^*$. For the most poorly performing models, RMSE$_U$ is generally larger than RMSE$_V$ (Fig. 5 – circles plot above triangles).

Using the model classifications (Table 2, Fig. 1) we can evaluate whether particular approaches result in clear improvements in performance. It should be noted that the options used by groups were not always their most complex (compare ‘capability’ with ‘VL92’ options used in Tables 2, 3). Two sets of statistics are used: RMSE and the mean bias error (MBE) for day and night (Fig. 6, 7) with results for each model shown as a point for each class and category (Table 2). The range, inter-quartile range (IQR), mean and median performance of the category within the class can be compared. Perfect performance would have a RMSE and MBE of 0 W m$^{-2}$. Given the relative magnitude of the MBE for night-time $Q_E$ (< 12 W m$^{-2}$), these results are not considered further here.

First, the method to represent vegetation (V class 1) is considered. Of the 18 models that have the ability to include vegetation as a separate tile (Vs, Table 2), five did not. Six additional models have integrated vegetation (Vi) within their urban surface. For the VL92 runs, a total of 14 models do not consider vegetation (Vn). The IQR of RMSE (bars on Fig 6) is smaller in the day time for $Q^*$, $Q_H$ and $\Delta Q_S$ when vegetation is included as a separate tile (Vs). In the daytime, not including vegetation (Vn) results in the largest RMSE medians ($Q_H$=181, $\Delta Q_S$=136, $Q^*$=59, $Q_E$=36 W m$^{-2}$) and MBE medians ($Q_H$=158 W m$^{-2}$, $\Delta Q_S$=-107, $Q^*$=42, $Q_E$=-28). For daytime $Q_H$ and $Q^*$, the ‘tiled’ approach (Vs) has the smallest RMSE (median=71 and 46 W m$^{-2}$, respectively) and MBE (median=18 and -14 W m$^{-2}$, respectively), whereas the integrated vegetation (Vi) has the lowest individual RMSE values for $Q^*$ and $\Delta Q_S$.

Examining the combination of model characteristics (Table 3) shows that for those that do not take into account vegetation, Vn, share only one common characteristic: their calculation of $\Delta Q_S$ via conduction or net radiation (class 8, Sc, Sn). However, many models that do include vegetation (Vs) also use this approach to heat conduction (Sc), so this is not likely to be a primary co-explanation. Interestingly, not including vegetation even in this area where there is very little, and where the measured $Q_E$ is small compared to the other fluxes, appears to impact the ability to model $Q^*$ and $Q_H$, with a resulting poor performance also in $\Delta Q_S$.

The VL92 site also has low $Q_F$. Most groups assumed it is negligible (ANn) with only seven groups explicitly including the flux (Table 2). Those that have considered it have taken a wide range of approaches but because of the small numbers they are grouped together into one class for analysis (ANm). Similarly, different temporal approaches to modeling $Q_F$ (Tf, Tv) are used but the small number of models per class means analysis is the same and so is not shown separately. In the daytime, median RMSE and MBE are smallest for all fluxes when $Q_F$ is ignored (ANn). This differs for night time fluxes however, where ANm models have the smallest RMSE and absolute MBE for all fluxes except $Q_E$. 

www.megapoli.info
Figure 4: Ranked RMSE (W m$^{-2}$) (a-d) and normalised Taylor diagrams (e-h) associated with each model for the whole time period. Each model is randomly assigned a number and symbol. The key for the symbols is shown in plot (c). Shown are: (a, e) net all wave radiation, (b, f) turbulent sensible heat, (c, g) latent heat, and (d, h) storage heat fluxes. The dotted line is the mean RMSE. The Taylor plots display the correlation coefficient in relation to the polar axis comparing hourly values, the normalised standard deviation in relation to the horizontal axis and the normalised RMSE in relation to the internal circular axes (Taylor, 2001). (N=312 h).
Figure 5: Ranked RMSE (W m⁻²) for net all wave radiation (a, b, c) and turbulent sensible heat flux (d, e, f) by time of day (see text) for day (a, d), night (b, e) and transition (c, f) time periods. Circles and triangles are RMSEₐ and RMSEₜ, respectively. The mean observed flux (W m⁻²) is for each period is given.
Figure 6: RMSE (W m⁻²) for each of the seven categories (see Table 2 for key) Q*, QH, QE and ΔQS for day (left/top) and night (right/bottom). Each dot is a model, the shaded bar shows the 25th and 75th percentile, the line indicates the median and the 'x' the mean. The maximum and minimum are indicated by the triangles. Note that the vertical scale varies between graphs. Mean observed fluxes for QE is 28.6 (day) and 3.9 W m⁻² (night) and for ΔQS is 191.7 (day) and -70.7 W m⁻² (night) (for Q* and QH see Fig. 5).
Figure 7: As for Fig. 6 but for the mean bias error (MBE) (W m⁻²). For daytime $Q_E$, the RMSE and MBE are best for Vi (median=27 and 3 W m⁻², respectively). At night for $Q_H$, the performance is poorest for those models that assume a separate tile (Vs) (median RMSE=19 W m⁻², MBE=17 W m⁻²) and best for Vi models (median RMSE=14 W m⁻², MBE=-2 W m⁻²). However, for $Q^*$, Vi and Vn have similar performance (median RMSE Vi=16, Vn=19, MBE: Vi=-11, Vn=-8 W m⁻²).
The model combinations in Table 3 show that those models which use an internal temperature (ANi) tend to have a fixed or variable temporal variation in $Q_F$ (class 3, Tf, Tv), an urban morphology that is multi-layered (L4-7), and a surface albedo/emissivity which has three or more facets (class 7, AEf).

The urban morphology (class 4, L) has a relatively large within class difference (range of median RMSE: 98 W m$^{-2}$ and MBE: 130 W m$^{-2}$) for the daytime $Q_H$. For both RMSE and MBE, there is no clear best performer of models across all fluxes (best for RMSE median: $Q^*_{L3}$, $Q_{H Lm}$ & L1, $Q_E=L3$, $\Delta Q_S=L1$; best for MBE median: $\Delta Q_S=L1$, $Q_{H Lm}$, $Q^*_{L1}$, $Q_E=L1$) (Fig. 6, 7). At night, multi-layer models (Lm=L4-L7) perform best for $Q^*$, $Q_H$ and $\Delta Q_S$ based on MBE (median $Q^*=-1$, $Q_H=-1$, $\Delta Q_S=6$ W m$^{-2}$). The urban morphology classes have few common characteristics, although all L1 models use a single reflection and a bulk albedo and emissivity. Additionally, and by definition, L3 and all Lm models have three facets for albedo and emissivity.

With respect to the categorization based on facets and orientation (FO class 5), the largest difference is for the simulation of daytime $Q_H$ (difference between category medians $\Delta$RMSE of 96 W m$^{-2}$, $\Delta$MBE=129 W m$^{-2}$). Those that treat the surface as a ‘whole’ (FO1) have the lowest daytime RMSE for $Q_H$ and $\Delta Q_S$ (although for $Q_H$, median RMSE for FO1, FOo and FOi differ by < 8 W m$^{-2}$ while it is lowest for FOo and $Q^*$ and for FOi and $Q_E$). At night, the lowest median RMSE is: $Q^*=FOo$, $Q_{H FO1}$ and FOo, $Q_E=FO1$ all groups equal, $\Delta Q_S=FOo$. There is no consistency in groupings with the smallest daytime MBEs ($Q^*=FOo$, $Q_{H FO1}$, $Q_E=FO1$, $\Delta Q_S=FOo$). Except for $Q^*$, during the daytime, models that simulate a canyon but have no associated orientation (FOo), have the largest biases ($Q_H$ - positive bias, $Q_E$ and $\Delta Q_S$ - negative bias) and these are likely to be complementary. At night, models that incorporate orientation and intersections (FOi) have the smallest bias, again except for $Q^*$, where it is FOo models (although differing by just 1 W m$^{-2}$ compared with FOi). In the daytime, for $Q_H$, the median RMSE performance improves from FOo, FOo, FOi, FO1 (165, 77, 74, 69 W m$^{-2}$, respectively) and for $Q^*$, improves from FOi, FOo, FO1, FOo (67, 52, 46, 43 W m$^{-2}$). The unique combinations that these categories of models have in common include those that treat the surface as a ‘whole’ (FO1), have no anthropogenic heat fluxes calculated (ANn) and obviously, have just a slab surface morphology, single reflections and a bulk albedo and emissivity. Models that include orientation (FOo, FOi) all assume three or more facets for albedo and emissivity (AEf) (as would be expected). Models without orientation (FOo) largely utilize conduction methods to calculate the storage heat flux (Sc).

When the models are classified based on the number of reflections used, there are large within class differences ($\Delta$RMSE= 89 W m$^{-2}$ for daytime $Q_H$) (Fig. 6). This is also the largest difference for the MBE ($\Delta$MBE=109 W m$^{-2}$) (Fig. 7). During the day, models with a single reflection scheme (class 6, R1) perform best for all fluxes except $Q_E$ (median RMSE $\Delta Q_S=98$, $Q_{11}=73$, $Q^*=46$ W m$^{-2}$). The daytime MBE is smallest for $Q^*$ models that calculate single reflections (Rs) (median MBE $Q^*=-14$). Generally, during the daytime the models which have infinite reflections (Ri) perform least well for $Q_H$ and $Q_E$, (median MBE $Q_{11}=147$, $Q_{E}=-27$ W m$^{-2}$; median RMSE $Q_{11}=162$, $Q_E=-35$ W m$^{-2}$); there are also negative median MBEs for all classes for $Q_E$ and $\Delta Q_S$, while $Q_{11}$ and $Q^*$ have a positive bias, with the exception of the single reflection class and $Q^*$. This suggests that the single reflection models may not allow enough radiation to be absorbed compared with observations. For $\Delta Q_S$ and $Q_H$, RMSE increases with the number of reflections modeled.

At night, models using increasing numbers of reflections have smaller RMSE for $Q^*$ ($Q^*: R1=13$, Rm=20, R1=28 W m$^{-2}$), whereas the situation reverses for $Q_H$ and $Q_E$, with those modeling fewer reflections yielding better results ($Q_H$: R1=27, Rm=18, R1=17 W m$^{-2}$). For the calculation of $Q^*$ at night the Ri type models perform best with the lowest median RMSE.
and MBE (RMSE=13, MBE=4 W m$^{-2}$). However as for daytime, superior performance for one flux is accompanied by poorer performance in another. All approaches have a similar sized negative MBE for nocturnal $\Delta Q_S$ (median=-21 to -22 W m$^{-2}$). The MBE for single reflections suggests that the surface temperature is too high, but correcting the bias during the daytime is likely to increase the nocturnal surface temperature, so there may be other issues with the models that use this method. Compensation also occurs between $Q^*$ and $Q_H$ most particularly at night. All schemes with infinite reflections (Ri) have three facets for albedo and emissivity (AEf).

The differences within groups of models are amongst the greatest when stratified based on specification of albedo/emissivity (class 7, AE). In general, using a bulk albedo/emissivity (AE1) results in better performance for all fluxes during the day based on median RMSE and MBE (median MBE $\Delta Q_S$=-23, $Q^*$=3 $Q_E$=10, $Q_H$=28 W m$^{-2}$). Models using two facets (AE2) tend to have the poorest daytime performance (except for $Q^*$ where median MBE for all groups is similar). At night, the differences in median MBE are smaller ($Q_H$: AE2=14, AEf=16 AE1=9; $Q^*$: AE1=14, AEf=13, AE2=17 W m$^{-2}$). In this evaluation, where buildings are small and widely spaced, the ability to distinguish different facet characteristics of albedo and emissivity is not important. However, where buildings are taller, more tightly spaced and/or with very contrasting materials, this result may not necessarily be the same. It is also important to remember that depending on the intended application, the ability to change facet material characteristics may be very important; for example for scenario testing (e.g. for urban heat island mitigation).

Classifying models based on method used to calculate $\Delta Q_S$ (S class 8) has a relatively small difference in the median RMSE and MBE for all fluxes. Again the biggest difference in performance is associated with daytime $Q_H$ (52 and 74 W m$^{-2}$ for RMSE and MBE). The daytime $Q^*$ differences are 6 and 12 W m$^{-2}$ for RMSE and MBE, respectively; these are the smallest within-group differences in median for $Q^*$ across the classes. The residual method (Sr) performs better during the daytime for all fluxes except $Q^*$ daytime (median MBE: $Q_H$=27, $Q^*$=6, $Q_E$=-11, $\Delta Q_S$ =-30 W m$^{-2}$) and for all night time fluxes (median MBE: $Q_H$=11, $Q_E$=4, $\Delta Q_S$ =-5, $Q^*$ =-5 W m$^{-2}$). Sc models often assume three facets (AEf) without orientation (Fo).

If the 31 different classes are considered, the best performance during the daytime for $Q^*$ are from the FOo class (median RMSE of 43 W m$^{-2}$). There are two classes with an absolute median MBE of $\leq$ 3 W m$^{-2}$ (L1, AE1). There are six models with both these characteristics (Table 3). For daytime $Q_H$, there are four classes with a MBE of $< 20$ W m$^{-2}$ (Vs, Lm, FO1, FOi). There is only one model with all of these (viz; Vs, Lm, FOi). The best overall performance for daytime $Q_H$, based on median RMSE, has a value of 69 W m$^{-2}$ (FO1), but there are seven other classes within 4 W m$^{-2}$ of this (Vs, Vi, Ls, Lm, R1, AE1, Sr) and three additional classes within 7 W m$^{-2}$ (ANm, FOo, FOi), thereby accounting for all seven major classes (Table 2). No models have all of these characteristics, while two have five of them but do not generally fall within the group of best performing models.

At night, best performance for $Q^*$ is associated with ANm, Lm, Ri (median RMSE 11-13 W m$^{-2}$ and/or median MBE $< |4|$ W m$^2$) and for $Q_H$ with Vi, Lm and FOi (median MBE$=-2$, -1 and -6, median RMSE=14, 27 and 27 W m$^{-2}$, respectively). Sr and Sc models have a similarly good RMSE (17-18 W m$^{-2}$).

For daytime $Q_E$, best overall performance is from Ls, FO1, AE1, Vi, FOi, R1 (median MBE $< |10|$ W m$^2$). For daytime $\Delta Q_S$, models with median RMSE $< 96$ W m$^2$ are Vi, Ls, FO1, AE1 and Sr but based on the absolute MBE, the best performing models are FOo (median MBE$=\leq |10|$ W m$^2$) and FO1, AE1, Sr, Vi, Lm, Ls (median MBE$< |30|$ W m$^2$). At night Lm, ANm, FOi, Sr and Vi models perform well based on median MBE and RMSE ($<|6|$ and/or $<22$ W m$^2$).
6. Conclusions

Urban surface-atmospheric exchanges are modelled for a wide variety of applications. The large set of models, examined here, have a range of approaches, complexities, and parameter requirements. Through the first stage of the first international model comparison reported here, significant model developments have taken place and improvements in model performance have resulted.

Evaluation of 33 models, with Vancouver (V L92) data, shows that generally models have best overall capability to model $Q^*$ and least capability to model $Q_E$ (order $Q^*$, $\Delta Q_S$, $Q_H$ and $Q_E$, Table 6). No model performs best or worst for all fluxes. In particular, it seems to be difficult to minimise both $Q^*$ and $Q_H$ errors. There is evidence that some classes of models perform better for individual fluxes but not overall. Typically, those that perform best during daytime do not perform best at night.

The daytime RMSE for $Q_H$ is larger than for $Q^*$ for all but four models. These four are characterised as having amongst the four largest $Q^*$ RMSE values. For RMSE$_S$, there is the tendency for $Q_H$ errors to be greater than for $Q^*$, although there are more cases where the errors are similar. The unsystematic errors are generally smaller than systematic errors, particularly for the most poorly performing models. For most models, $Q_H$ has a positive MBE which observational errors may contribute to.

Seven characteristics (relating to: vegetation, $Q_F$, morphology, facets and orientations, reflection, albedo and emissivity, $\Delta Q_S$) are used to classify each model. Some of the greatest differences in model performance are found between classes of model that treat vegetation and reflections differently. Some of the smallest differences relate to approaches used to calculate the $\Delta Q_S$ followed by urban morphology. Not including vegetation, even at a site with limited vegetation, results in the poorest performance for all fluxes during the day (in terms of RMSE) and for $Q_E$ at night. During the day, median RMSE for models that do not include $Q_F$ is similar (or better) than for those that do. However, at night median RMSE for models which include $Q_F$ shows better performance for $Q^*$, $Q_H$ and $\Delta Q_S$. Models which account for urban morphology orientation, and also intersections, often have slightly better performance than schemes which do not (e.g. $Q_H$ in the day time). The addition of intersections, however, does not always improve performance appreciably and in some cases has a negative impact on model performance.

The results for reflection schemes vary between day and night and with statistical measure (RMSE or MBE). In general, using a bulk albedo/emissivity results in better performance for all fluxes during the day. Classifying based on method used to calculate $\Delta Q_S$ has the smallest difference in the median of the RMSE and MBE of all classes. The residual method performs better during the day for all fluxes while at night, differences are less significant. Class combinations show no models display all characteristics associated with strongest performance within the various, although two display a large proportion of these. In general, the simpler models perform as well as the more complex models based on all statistical measures.

These results are based on a short time series for one urban location. In Phase 2, the same models will be evaluated using a second dataset (Grimmond et al. 2009b). These results raise a number of questions which will be considered, with different flux partitioning, a wider range of conditions, and a longer time series. Of particular interest is whether the same models and classes perform well; whether the relative ability to model the individual fluxes remain the same; and whether it is possible for any class of models to minimize errors in both $Q^*$ and $Q_H$. 
Acknowledgements

Funded (Grimmond) by the Met Office (P001550), European Union Framework 7 (7 FP7-ENV-2007-1): MEGAPOLI (212520), BRIDGE (211345). This is a contribution to COST728 (Enhancing mesoscale meteorological modeling capabilities for air pollution and dispersion applications). We thank all involved in the original dataset collection (especially James Voogt and Tim Oke) and their funding agencies. Funding from BSIK-COM29 (Steeneveld), CATER 2006-2202 (Baik), and other agencies that support all the time contributed by participating groups are acknowledged. Colour versions of the figures can be obtained by emailing Sue Grimmond (sue.grimmond@kcl.ac.uk).

References

Best, M.J., 2005: Representing urban areas within operational numerical weather prediction models, Bound.-Layer Meteor., 114, 91-109.
Fortuniak, K., B. Offerle, and C.S.B. Grimmond, 2004: Slab surface energy balance scheme and its application to parameterisation of the energy fluxes on urban areas, NATO ASI, Kiev, Ukraine,


Part III
Initial Results from Phase 2 of the International Urban Energy Balance Comparison Project

CSB Grimmond1†, M Blackett1, MJ Best2, J-J Baik3, SE Belcher4, J Beringer5, SI Bohnenstengel4, I Calmet6, F Chen7, A Coutts8, A Dandou8, K Fortuniak9, ML Gouveia1, R Hamdi10, M Hendry2, M Kanda11, T Kawai12, Y Kawamoto13, H Kondō14, ES Krayenhoff15, S-H Lee3, T Loridan1, A Martilli16, V Masson17, S Miao18, K Oleson7, R Ooka13, G Pigeon17, A Porson2,4, Y-H Ryu3, F Salamanca16, GJ Steeneveld19, M Tombrou8, JA Voogt20, D Young1, N Zhang21

1. Introduction

Land surface schemes (LSS) model energy exchanges between the surface and the atmosphere for a wide range of different environments (e.g. deciduous trees, coniferous trees, grasses, bare soil, urban). They provide the lower boundary conditions (fluxes) to meso- and global scale atmospheric models and are forced with data from the overlying model. A wide variety of approaches are taken to model the influence of the underlying land-surface type. To model the exchanges for an urban environment, LSS range from a relatively simple representation of the urban environment as an impervious slab, to models which take into account the 3-d geometry of buildings with varying heights and material characteristics (Grimmond et al. 2009, 2010). During the process of simplification inherent to modelling, urban LSS (ULSS) 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 of the number of parameter 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 ULSS have been evaluated individually against observational datasets (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, Porson et al. 2009, Loridan et al. 2010). Although providing useful insights, these studies lack a structure that facilitates robust inter-comparison. 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 ULSS in a common and consistent manner. In the first paper (Grimmond et al. 2010), a short data set (14 days) was provided to participants for a known site. The dataset consisted of both the forcing data and the observed flux data. Here the results from a comparison of 32 urban LSS (Table 1), which represent a range of approaches (Figure 1), are analysed for a longer data set (16 months) with the participants initially not knowing the location of the site beyond that it is urban. All participants in the second phase had to have completed Phase 1 (Grimmond et al. 2010). There is one model from Phase 1 that did not participate in 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. While each group is informed how their own model
performs; the anonymity of other models’ results is maintained.

The objectives of this paper are:

1. To evaluate the ability of ULSS, 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 areas for future research and improvement for the models.

The first objective aims to highlight what might be expected in terms of ULSS 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.

Table 1: The number of versions of each model used in the comparison and number of groups using it.

<table>
<thead>
<tr>
<th>Code</th>
<th>Model Name</th>
<th>References</th>
<th>Versions</th>
<th>Groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>BEP02</td>
<td>Building Effect Parameterization</td>
<td>Martilli et al. (2002)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>BEP_BEM08</td>
<td>BEP coupled with Building Energy Model</td>
<td>Martilli et al. (2002), Salamanca et al. (2009), Salamanca and Martilli (2009)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CLMU</td>
<td>Community Land Model – Urban</td>
<td>Oleson et al. (2008a, b)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>IISUCM</td>
<td>Institute of Industrial Science Urban Canopy Model</td>
<td>Kawamoto and Oka (2006; 2009a; b)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>JULES</td>
<td>Joint UK Land Environment Simulator</td>
<td>Essery et al. (2003), Best (2005), Best et al. (2006)</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>LUMPS</td>
<td>Local-scale Urban Meteorological Parameterization Scheme</td>
<td>Grimmond and Oke (2002), Offerle et al. (2003), Loridan et al. (submitted)</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>NKUA</td>
<td>University of Athens Model</td>
<td>Dandou et al. (2005)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>MORUSES</td>
<td>Met Office Reading Urban Surface Exchange Scheme</td>
<td>Harman et al. (2004 a,b), Porson et al. (submitted)</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>MUCM</td>
<td>Multi-layer Urban Canopy Model</td>
<td>Kondo and Liu (1998), Kondo et al. (2005)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NJU-UCM-S</td>
<td>Nanjing University Urban Canopy Model-single layer</td>
<td>Masson(2000), Kusaka et al. (2001)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>NJUC-UM-M</td>
<td>Nanjing University Urban Canopy Model-multiple layer</td>
<td>Kondo et al.(2005), Kanda(2005a; b)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SM2U</td>
<td>Soil Model for Submesoscales (Urbanized)</td>
<td>Dupont and Mestayer (2006), Dupont et al. (2006)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SNUUCM</td>
<td>Seoul National University Urban Canopy Model</td>
<td>Ryu et al. (2009)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SRUM2/SRUM4</td>
<td>Single Column Reading Urban Model tile version</td>
<td>Harman and Belcher (2006)</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>SUEB</td>
<td>Slab Urban Energy Balance Model</td>
<td>Fortuniak (2003), Fortuniak et al. (2004, 2005)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SUMM</td>
<td>SUMM (Simple Urban Energy Balance Model for Mesoscale Simulation)</td>
<td>Kanda et al. (2005a,b), Kawai et al. (2007, 2009)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TEB</td>
<td>Town Energy Balance</td>
<td>Masson (2000), Masson et al. (2002), Lemonsu et al. (2004), Pigeon et al. (2008)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TEB-ml</td>
<td>Town Energy Balance with multilayer option</td>
<td>Hamdi and Masson (2008), Masson and Seity (2009)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TUF2D</td>
<td>Temperatures of Urban Facets 2D</td>
<td>Krayenhoff and Voogt (2007)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>TUF3D</td>
<td>Temperatures of Urban Facets 3D</td>
<td>Krayenhoff and Voogt (2007)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>VUCM</td>
<td>Vegetated Urban Canopy Model</td>
<td>Lee and Park (2008)</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

2. Methodology

To participate in this comparison the models had to simulate urban energy balance fluxes from the forcing data provided (Table 2). The urban energy balance for these purposes is defined as:

\[
Q^* + Q_F = Q_H + Q_E + \Delta Q_S
\]  

(1)

where \(Q^*\) is the net all wave radiation flux density which consists of the incoming short wave (\(K_\downarrow\)) and long wave (\(L_\downarrow\)) radiation, which was provided as part of the forcing data, and the
The outgoing short- \((K↑)\) and long-wave \((L↑)\) radiation which have to be modelled:

\[
Q^* = (K↓ - K↑) + (L↓ - L↑)
\]  

(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 \((\Delta Q_S)\). Advection 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 model. Here the ULSS are run independent of any large scale model (i.e. ‘offline’). This is to ensure that the model performance evaluates the ULSS 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 only discuss the results for the directly observed fluxes so the storage heat flux and anthropogenic heat flux are not discussed. These will be discussed in later papers.

To conduct this comparison, the principles of the PILPS are employed, whereby a set of forcing data is distributed to participants to run their models, with limited information about site characteristics beyond the designation urban. At the completion of each of the four stages, additional site information was provided (Table 2). In Stage 1 only the forcing data was provided along with knowledge that observations were for an urban area measured at 6.25 times the mean roughness height \((z_H)\). In later stages, more site information was provided, consisting of basic surface cover fractions (Stage 2); urban morphology (Stage 3); and characteristics of urban materials (Stage 4). From this information, further parameters could be derived by participants as necessary (Grimmond et al., 2010). After the completion of each run, participants provided their 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 data to allow seasonality to be incorporated into the modelling; (2) not been used extensively by modelling groups previously to test models; (3) an almost complete quality controlled data set available (i.e. not being available for certain meteorological conditions only); and (4) cooperation with those that were involved in the data collection to participate in PILPS-urban. The Phase 2 observation site was in suburban Melbourne, Australia (Coutts et al. 2007a, b). This location was concealed from participants until the completion of Stage 4 before which an equivalent latitude and longitude for solar zenith angle was released. The radiative fluxes were measured using Kipp & Zonen CM 7B and CG4 radiometers. Temperature and relative humidity were measured using a Campbell Scientific 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 (2). The turbulent sensible and latent heat fluxes were measured using the eddy covariance technique. A Campbell Scientific Inc. (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 analyzer (February 2004 to November 2004). They were sampled at 10 Hz and block averaged using a Campbell Scientific CR23X datalogger. The fluxes were calculated for 30 minute intervals (Coutts et al. 2007a,b). Diurnal and seasonal anthropogenic heat fluxes were estimated for the site, following Sailor and Lu (2004), includes sources of anthropogenic heat 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 (1). This has the inherent problem that it accumulates all the measurement errors and missing terms (e.g.
Table 2: Data provided at each stage. The exact latitude and longitude (*) were not known only an equivalent for solar zenith angle. The material characteristics provided at Stage 4 consisted of information for four layers for each facet (roof, wall and road) consisted of: 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 http://www.engineeringtoolbox.com/spesific-heat-capacity-gases-d_159.html d http://www.engineeringtoolbox.com/thermal-conductivity-d_429.html

<table>
<thead>
<tr>
<th>Category</th>
<th>Data provided</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stage 1</strong></td>
<td></td>
</tr>
<tr>
<td>Forcing data</td>
<td>K_air, L_air, air temperature, station pressure, specific humidity, wind components, rainfall</td>
</tr>
<tr>
<td>Site</td>
<td>Latitude*, Longitude* Measurement height: 6.25 mean roughness height</td>
</tr>
<tr>
<td><strong>Stage 2</strong></td>
<td></td>
</tr>
<tr>
<td>Plan area fraction</td>
<td>Pervious = 0.62 Impervious=0.38</td>
</tr>
<tr>
<td><strong>Stage 3</strong></td>
<td></td>
</tr>
<tr>
<td>Plan area fraction</td>
<td>Surface cover Fraction Total</td>
</tr>
<tr>
<td>Building</td>
<td>0.445 Impervious</td>
</tr>
<tr>
<td>Concrete</td>
<td>0.045 0.62</td>
</tr>
<tr>
<td>Road</td>
<td>0.130</td>
</tr>
<tr>
<td>Vegetation (excl. grass)</td>
<td>0.225 Pervious 0.150 Gras 0.005 Other (bare or pools)</td>
</tr>
<tr>
<td>Other</td>
<td>Urban climate zone=5 Population density= 415.78 km⁻²</td>
</tr>
</tbody>
</table>
(b) Stage 4: details of layers components for each facet

<table>
<thead>
<tr>
<th>Wall</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Material</td>
<td>%</td>
<td>$C_p$</td>
<td>$c^a$</td>
<td>$\lambda$</td>
<td>d</td>
<td>Material</td>
<td>$C_p$</td>
<td>$c^a$</td>
<td>$\lambda$</td>
<td>d</td>
<td>Material</td>
<td>$C_p$</td>
<td>$c^a$</td>
<td>$\lambda$</td>
</tr>
<tr>
<td>Brick</td>
<td>27.94</td>
<td>840.0</td>
<td>1.26</td>
<td>0.71</td>
<td>110</td>
<td>Metal</td>
<td>30</td>
<td>1105</td>
<td>8.73</td>
<td>72.00</td>
<td>Asphalt</td>
<td>75</td>
<td>920</td>
<td>1.10</td>
</tr>
<tr>
<td>Softwood</td>
<td>59.43</td>
<td>1975.5</td>
<td>1.11</td>
<td>0.14</td>
<td>20</td>
<td>Concrete</td>
<td>40</td>
<td>837</td>
<td>1.52</td>
<td>1.10</td>
<td>Concrete</td>
<td>25</td>
<td>837.0</td>
<td>1.5</td>
</tr>
<tr>
<td>Concrete</td>
<td>5.37</td>
<td>837.0</td>
<td>1.52</td>
<td>0.87</td>
<td>100</td>
<td>Terracotta</td>
<td>20</td>
<td>837</td>
<td>1.42</td>
<td>0.99</td>
<td>Asbestos</td>
<td>35</td>
<td>72.00</td>
<td>3.0</td>
</tr>
<tr>
<td>Asbestos cement</td>
<td>5.37</td>
<td>1005.0</td>
<td>1.93</td>
<td>0.58</td>
<td>8</td>
<td>Asbestos cement</td>
<td>10</td>
<td>945</td>
<td>1.79</td>
<td>0.55</td>
<td>Coarse crushed rock</td>
<td>80</td>
<td>1004</td>
<td>1.56</td>
</tr>
<tr>
<td>Concrete/wood</td>
<td>1.13</td>
<td>1406.2</td>
<td>1.31</td>
<td>0.79</td>
<td>60</td>
<td>Tile</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Metal</td>
<td>0.76</td>
<td>1105.0</td>
<td>8.73</td>
<td>72.00</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 Insulation (air)</td>
<td>80</td>
<td>1880</td>
<td>1.50</td>
<td>0.23</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Road</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Material</td>
<td>%</td>
<td>$C_p$</td>
<td>$c^a$</td>
<td>$\lambda$</td>
<td>d</td>
<td>Material</td>
<td>$C_p$</td>
<td>$c^a$</td>
<td>$\lambda$</td>
<td>d</td>
<td>Material</td>
<td>$C_p$</td>
<td>$c^a$</td>
<td>$\lambda$</td>
</tr>
<tr>
<td>Air</td>
<td>85</td>
<td>1010</td>
<td>0.0013</td>
<td>0.024</td>
<td>50</td>
<td>Coarse crushed rock</td>
<td>100</td>
<td>712</td>
<td>0.04</td>
<td>0.03</td>
<td>Soil (sandy loam)</td>
<td>80</td>
<td>1004</td>
<td>1.56</td>
</tr>
<tr>
<td>Brick</td>
<td>20</td>
<td>840.0</td>
<td>1.26</td>
<td>0.71</td>
<td>110</td>
<td>Fine crushed rock</td>
<td>20</td>
<td>840.0</td>
<td>1.26</td>
<td>0.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gypsum / plaster board</td>
<td>0.75</td>
<td>712</td>
<td>0.04</td>
<td>0.03</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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).
horizontal advection $\Delta Q_\alpha$ in this flux (Grimmond and Oke 1999, Offerle et al. 2005). However, Offerle et al. (2005) and Roberts et al. (2006) obtain similar results when detailed facet temperature measurements are used to determine this flux compared to local-scale residual estimates of $\Delta Q_\alpha$. For all observations there are measurement errors so the observed fluxes and the forcing data are not without errors.

The forcing data consisted of 22,772 continuous 30-minute intervals (474.5 days) from August 2003 – November 2004. Not all of the fluxes were available during all of these intervals, so here analysis is limited to the periods where 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 initialisation, period (the impact of this will be evaluated in a future paper). The post-initialisation period allows for performance through an annual cycle to be evaluated.

Here, 32 different ULSS are compared (Table 1). 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) which are discussed more fully 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 categorised according to complexity (either simple or complex) (Figure 1). Models were further categorised by their overall complexity dependent on the number of ‘complex’ or ‘simple’ characteristics they possessed. The three groups are: (a) ‘complex’ when all characteristics were complex, (b) ‘medium’ when the models possess one or two simple characteristics; and (c) ‘simple’ when they had three or more simple characteristics. Vegetation is not incorporated into this classification.

Comparison statistics reported here include: root mean square error (RMSE), with both systematic (RMSEs) and unsystematic (RMSEu) components, and the mean bias error (MBE) and the coefficient of determination ($r^2$). These are formally defined in Grimmond et al. (2010). A larger systematic error, typically indicates that the model has a problem in the model physics or parameter values, whereas a large RMSEu should be associated with the ability 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 models’ ability to simulate radiative fluxes the first aspect considered is whether there is closure in the radiation budget. This is 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, 15/32 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 10 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 a $r^2$ of 0.999999), then Stages 1/2/3/4 have 16/14/16/13 models, respectively. These 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 in Stages 1 and 2, which jump to greater than 0.998 at Stage 3.

Each modelling group which had a case of non-closure was asked to determine the cause. The models without radiation balance closure problems are classified as P0 in the following analysis. The explanations for those that did not have closure include (classified in analysis): not using the provided forcing data (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: rather than using the individual 30 min interval forcing $K_{↓}$ data, the daily peak observed $K_{↓}$ was used instead and the other time periods for the day were obtained by assuming clear sky conditions, resulting in overpredicted $K_{↓}$ and therefore $Q^{*}$ (4 cases, P1); and, the observed $L_{↓}$ data were not used but modelled (1 case, P1).

In the second situation, fluxes calculated independently, the ULSS calculate $Q^{*}$ but for the purpose of this comparison, the radiative components have been calculated (3 cases, P1) or there is an additional term in $L_{↑}$ which is not incorporated into $Q^{*}$ (1 case, P4). In the third situation, which relates to timing, the lack of closure is related to either 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 (2 cases, P3). With an explicit temporal numerical model for the shortwave: $K_{↑}(t) = αK_{↓}(t-δt)$ where $δt$ is the time step of the model (e.g. 300s). This requires the forcing data to be interpolated which for $K_{↓}$ may be questionable. For $L_{↑}$, this depends from an emitted contribution from the surface temperature and a reflected part: $L_{↑}(t) = (1-ε)L_{↓}(t-δt)+εσT_{S}^{4}(t)$. The surface temperature $T_{S}$ depends of the energy received and has inertia. Alternatively it is because $K_{↑}$ is only calculated if the sun is above the horizon for the whole time interval (1 case, P3), thereby impacting the day length. For the fifth situation, spatial resolution (2 cases, P3), in the process of rasterizing the surface causes the total sky view factor of all the model patches to not add up to 1.0. This means that the models absorb slightly too much or too little diffuse solar or longwave radiation. The final situation is where there are some problems which the modelling groups have not been able to determine, leading to the imbalance (3 cases, P4).

### 3.2 Outgoing shortwave radiation

The performance of each model, with respect to outgoing shortwave radiation ($K_{↑}$), is shown in Figure 2 based on their RMSE, with the models which do not have closure indicated. For this upwelling solar flux, only daytime fluxes are analysed giving 4266 - 30 min periods for comparison. For these data the mean observed flux is 54.2 W m\(^{-2}\). At Stage 1 for $K_{↑}$ the mean RMSE for all (N=32)/(N=31 models)/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 to 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 2). This would be expected to result in notable improvements as albedos for appropriate fractions (‘urban’, ‘vegetated’) could be assigned more realistically. However RMSE for five models became poorer. In Stages 2 to 3 a total of 14 models reduced (and 14 models increased) their RMSE and 13 in Stage 3 to 4 (and four 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 about the wall heights. In the pervious fraction, grass could be distinguished from other vegetation. As expected, at Stage 4 when the site observed albedo was provided, there was the largest overall improvement in $K_\uparrow$ based on the mean and median RMSE (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; for example, 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 that the user’s choice of parameter values 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). This suggests 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) between simple (s, yellow, light grey), medium (m, blue, medium grey), and complex (c, purple, 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, but the model that has dropped out and a Cc model. At Stage 4, the majority of the Cs models are doing well but the poorest performing model belongs to that group.

The effective albedo ($\alpha_{\text{eff}}$) used in the models can be determined from $K_\uparrow_{\text{mod}}/K_{\text{obs}}$. 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 $\alpha_{\text{eff}}$ 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 to 0.20); and 16 cases with an $\alpha_{\text{eff}}$ within 0.125-0.175, of which 11 have the lowest RMSE for $K_\uparrow$. For June 21 there is a similar distribution. The slightly higher $\alpha_{\text{eff}}$ (0.175-0.18) values are associated with the next best cohort of RMSE performance.

The average cohort mean bias error (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 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.

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 2) 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.
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, purple, dark grey) (see section 2 for definition). The mean observed flux for this period was 54.2 W m$^{-2}$. The lower row (e-j) shows the RMSE for the classes by approach taken (see Figure 1 for code interpretation or text). Individual models are shown by the points, maximum and minimum by the triangles and the inter-quartile range 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.
For example Figure 2e 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) which had 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 2e). By Stage 3 the Vi models have a median RMSE of less than 4 W m^{-2} which is the smallest. From Stage 2, when more models included vegetation (Vs models increase in number at the expense of Vn) the model cohorts retained the same ordering Vn, Vs, Vi (decreasing median RMSE) but both Vn and Vs median performance is deteriorated 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).

The approach to specifying the urban morphology (L) is characterised by seven different approaches; from a slab surface (L1) to single layer models (L2 - two components, L3 - three facets) and multi-layer (L4-7). 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 2f). 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 not a 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 2g). 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 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 infinite 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 2h). By Stage 4 the median RMSE of 6 W m^{-2}, which 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}). This is however not the case for the Ri group (median RMSE= 17 W m^{-2}) which 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. Significant improvement for these models in 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 models continue to have a wide range of values for the individual models.

The models that do not have problem with net radiation balance closure (P0) have the smallest median RMSE at each Stage (Figure 2i). 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. The P1 models which have problems calculating a component of the radiative balance or did not use the forcing data for individual time intervals perform poorly throughout.

The models have been classified based on their overall complexity (Figure 1) as simple (Cs), medium (Cm) and most complex (Cc). This is simply a function of the number of complex attributes used within the models (see section 2). For all three approaches there are steady improvements in performance as additional information is provided (Figure 2j). 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 the models generally model $K_\uparrow$ 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 which 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.

### 3.3 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 daytime 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 to 4 are 16/14/14/17 W m$^{-2}$ respectively (Figure 3). 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 2), 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. From Stage 3 to 4 was also when the largest individual performance deterioration occurred (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 3) 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\(_\uparrow\). 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 performing model which has a larger RMSE\(_U\) and a small positive MBE.

The models that have radiative close (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 that is not 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.

In the different classifications of the models there is no clear approach which is better than the others. In most cases the model cohorts show poorest performance for all classes in Stage 4. For example, at Stage 4 the IQR is greater than in Stage 3 for 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 (Figure 3f). 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 reflections (R), and albedo and emissivity (AE), similar results are obtained.

The models perform generally better at night than over the 24 h period (mean observed flux day = 410. 14, night=368.98 W m\(^{-2}\)). At night, the median RMSE for Stages 1 to 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 to 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
Figure 3: As for Figure 2 but for outgoing longwave radiation ($L_\uparrow$) for all hours. The mean observed flux for this period was 389.6 W m$^{-2}$. Note lower plots are cut off at 0.40 of the maximum.
best performing (median RMSE W m$^{-2}$) models have Vi(16)/L2 (17)/FOi (18)/Ri (15)/AE1 &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 largest P class (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 3/2/3 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 the wide range of materials that are in urban areas and the difficulty of deciding what the appropriate values should be given the wide range of values that are found for many materials this may suggest that until there is away to obtain realistic values for actual sites that this may not be worth the effort to obtain the information. Here we contacted a large number of people associated with the building and planning design plus materials suppliers (see acknowledgements).

3.4 Net all wave radiation

Figure 4 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 2-4 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 worse 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 (Figure 2, 3). 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 there is a clear re-ranking between fluxes. Notably those that perform poorly for an individual component flux are not the poorest for $Q^{*}$ (Figure 2-4). This means that the application that the model is being used for is important; for example, assessing a mitigation strategy’s impact (such as changing the albedo of the materials) using an ULSS may be modelling the most directly impacted flux well, but not able to model the other fluxes well, 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 4). 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 $K_{\uparrow}$ and $L_{\uparrow}$. The models generally have a negative MBE (Figure 4, Stage 4 median -6 W m$^{-2}$). The models with the largest absolute MBE are both positive and negative (Figure 4).

The best and poorest performing model at all stages are medium complexity (Cm) models. 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 (remember model 17 no longer appears), with them having three out of the first six. The Cm are more grouped at the end with poor performance.

The models that do not account for vegetation (Vn) show a steady decline in performance across all stages but there is no strong evidence 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$^{-2}$. The same median RMSE is also the best when the models are sorted by their approach to facets and orientation (FO) for the simplest (FO1) at Stage 2, although FOo is only slightly larger at the same stage. This is repeated again for classification based on treatment of reflections (R1) and for albedo/emissivites (AE1).

The models with radiative closure (P0) have their lowest median RMSE at Stage 2 (15 W m$^{-2}$), which is the lowest cohort using this classification, and their largest at Stage 4 (25 W m$^{-2}$). The lowest median RMSE for P1 models is Stage 3 but these models have the largest IQR in Stages 3 and 4 (Figure 4i). Like $L_{\uparrow}$ 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 and having 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 to 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 to 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. For example the Cm models have daytime median RMSE of 50 W m$^{-2}$ at Stage 4.

Overall the simpler characteristics are often the best performing 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.
Figure 4: 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$^{-2}$. Note lower plots are cut off at 0.50 of the maximum.
3.5 Turbulent sensible heat flux

The errors are larger for the turbulent sensible heat flux (\(Q_H\)) than for the radiative fluxes (compare Figures 2-5). As for the radiative fluxes the provision of information about the fraction of vegetation (Table 2) resulted in an improvement (reduction in median RMSE) from 62 to 55 W m\(^{-2}\) (32 models). There was a similar sized reduction, down to 49 W m\(^{-2}\), at Stage 3 but at Stage 4 there is a small decline 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 in the measurements of 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 that reduce their RMSE at each stage (18 improved from Stage 1 to 2; and 10 which improved from Stage 2 to 3). There are also models whose performance deteriorates between stages; for example, 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 have 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 2\(^{nd}\) best at Stage 4) is the model which did best for the K\(_↑\), although it did not do best for Q* or L\(_↑\). 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 benefitting from additional information. However, there is improvement within the middle, most notably model 16 which performs best in Stage 4. The behaviour of the individual models with respect to systematic error show 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 5).

The models without radiation balance closure problems (P0) have a lower median RMSE than those that do not close (P1, P3, P4) except for at Stage 4 (P4) (Figure 5i). This is also 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 anthropogenic heat flux enters their 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. At Stage 1 there are 15 models that have closure, only one of these has evidence of a larger error in the model. As we currently cannot separate the role of anthropogenic heat in energy balance closure this will be investigated in later studies.
Figure 5: As for Figure 2 but for turbulent sensible heat flux (QH) for all hours. The mean observed flux for this period was 37.9 W m$^{-2}$. Note lower plots are cut off at 0.90 of the maximum.
The impact of treatment of vegetation is seen clearly when comparing the Vn models to the Vs and Vi. The Vn models have the widest range, largest IQR and the poorest median performance (Figure 5e). 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. Hence, more complex and realistic treatments of vegetation may be important for QH.

The simplest models with respect to morphology (L1) perform best relative to the others and improve across the stages (Figure 5f). 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 5g). The treatment of surface temperature (Figure 1) for the built 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 anthropogenic heat (AN) varies from not including or assuming it is negligible (ANn) to prescribing a value (ANp) to modelling it explicitly or 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 5j). 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 they 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 QH 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 Stage 1/4) and negative at night by Stage 4 (10/-8 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\(^{-2}\). 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 daytime 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 where for the more complex models this is not the case. In this case, where anthropogenic heat flux is not very large, the models that do not account for anthropogenic heat flux do better. The slab or bulk models also show a consistent improvement at each stage.

### 3.6 Turbulent latent heat flux

The modelling of latent heat flux (Q_E) needs to deal with the loss of water from a wet surface, for example after rainfall from roofs, roads and vegetation; and the transpiration of vegetation which continues between rainfall events. The median RMSE for the modelled
latent heat flux (Figure 6) dropped by the largest amount at Stage 2 when the 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 non-grass, Stage 3). Across the four stages there are seven models that have a large RMSE\(_{S}\) (58 W m\(^{-2}\)) and a 0 W m\(^{-2}\) RMSE\(_{U}\) so they are ignoring latent heat flux completely. There are a couple of models that are addressing some aspect of this flux but have even poorer performance than those 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; eleven 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 at Stage 2 and retains across the stages.

Overall the systematic errors are generally larger than the unsystematic errors. This is largely due to the models not attempting to model latent heat flux (Figure 6). 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 it. 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 6). However at Stage 2, when vegetation fraction came known, Cm models start to improve. By Stage 4, we have all model type represented at the poor end but the five models with the lowest RMSE are Cs.

From Stage 1 to Stage 2 three more models chose to include vegetation (Figure 6e). The three models which incorporated vegetation did so by using separate vegetation tile(s) (Vs). The Vs approach is the most common approach used. Using this approach there was a 10 W m\(^{-2}\) improvement between Stages 1 and 2. This is because the separate tiles can now 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 fluxes is coming exclusively from the vegetation scheme that has been “coupled” to the urban scheme. These 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 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 6f). 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 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).
Figure 6: As for Figure 2 but for turbulent latent heat flux ($Q_L$) for all hours. The mean observed flux for this period was 32.5 W m$^{-2}$. Note lower plots are cut off at 0.80 of the maximum.
The models with radiative closure (P0) have a large 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 large 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 6j).

The daytime RMSE are larger than night-time period (Stage 1 median 71.21 W m\(^{-2}\)) and all hour which is when the observed flux is larger and more variable (mean observed flux day= 56.41, night =8.53 W m\(^{-2}\)). The night-time fluxes do not show any improvement in skill 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\(^{-2}\); 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 the performance. However, despite in Stage 4 knowing the site location, many models did a poorer job at this stage than they had done previously. The simpler models generally performed best but did not improve the most across the four 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 this flux appropriately. The models that use the separate tile scheme have the about the same overall performance as those that take an integrated approach. But there is much wider range of results from the separate tile models. This suggests that using vegetation schemes that have been tested in non-urban areas is 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 if the modellers assumed an additional water, such as irrigation, to be available for evaporation.

4. Conclusions

Groups around the world have run their urban land surface schemes in offline mode for four stages, where increasing in formation 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. In the process of running the models there were a small number of models that improved for an individual flux at each stage as new information was provided. Across the fluxes considered there are two or more models that have improvements at all stages for a particular flux (2, 2, 5, 4; K\(\uparrow\), L\(\uparrow\), Q*, Q\(\text{H}\), Q\(\text{E}\) respectively). However, there are other models that have a drop in performance, and cases where there is a systematic decline at all stages (0, 1, 4, 0, 1; K\(\uparrow\), L\(\uparrow\), Q*, Q\(\text{H}\), Q\(\text{E}\) respectively).

From the analysis of the data returned from the modelling groups in relation to the observed 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 has very significant implications for the application of any model. This 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 get the sensible heat flux right, 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. Data provided at Stage 2 usually had the largest impact on model performance. This
confirms the conclusions of Grimmond et al. (2010). Moreover, the fact that the RMSE for the latent heat flux is of the same order of the latent heat flux itself, may indicate that work needs to be done to improve the link between the built part and the vegetated part.

- **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 there is a clear re-ranking between fluxes. Notably those that perform poorly for an individual component flux are not the poorest for $Q^*$ (Figure 2-4). This means it is important that when a user applies a model they are aware of the performance of the ULSS not only the initial flux of interest but also for the other fluxes which the user may wish to infer impact. Given the increasing use of ULSS for assessing mitigation and adaptation strategies this is very important.

- **Overall the ULSS generally model $K_\uparrow$ well and additional surface information did 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 minimise the errors in the outgoing longwave radiation change with the time of day. Generally performance improved when the pervious/impervious fraction is provided but did not when heights and further surface fractions details were. Most models deteriorated when building materials details were available typically back to Stage 1 performance but in many cases even poorer. Given the difficulty to gather appropriate values this suggests that this may not currently be worth the effort to obtain the information. Or there is a need to ensure that the data are of much better quality that is currently ‘easily’ obtainable.

- **In general the radiative fluxes are modelled better than the turbulent fluxes** There is clear trade off in performance between net all wave radiation ($Q^*$) and turbulent sensible heat flux ($Q_{sh}$).

- 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, the characteristics $L1/OF1/R1/AE1/Cs$ 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 user’s 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 whereas that was not the case for the more complex models. 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 aggregations of processes and/or scales in more complex models (e.g., 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 better the biophysical interactions between the atmosphere and urban surfaces. Although the ability to 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. Whereas, 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 this study (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 the 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 2) benefited a large number of models. The albedo and emissivity 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. Hopefully additional analyses will shed more light on this.

These results are the first of a number of different studies that will be undertaken from these data. Future analyses will consider for example: the role of seasonality on performance; the range of parameter values that are used; the determination of optimized parameters; and the individual and groups will also analyse what they have learnt from participating in the model comparison. Our initial message is one of caution in applying ULSS because in general, no model performs well across all fluxes but is essential to account for vegetation. A final caveat, is that the observations also have errors which vary with time of day and season.

References
Dandou, A., M. Tombrou, E. Akylas, N. Soulakellis and E. Bossioli, 2005: Development and


http://www.engineeringtoolbox.com/thermal-conductivity-d_429.html


Fortuniak, K., B. Offerle, and C.S.B. Grimmond, 2004: Slab surface energy balance scheme and its application to parameterisation of the energy fluxes on urban areas, NATO ASI, Kiev, Ukraine, 4-15.05 2004, 82-83, [www.met.rdg.ac.uk/urb_met/NATO_ASI/talks.html].


Kanda, M., T. Kawai, M. Kanega, R. Moriwaki, K. Narita, and A. Hagishima, 2005a: A simple
Kawamoto, Y. and R. Ooka, 2006: Analysis of the radiation field at pedestrian level using a meso-scale meteorological model incorporating the urban canopy model, ICUC-6, Göteborg, Sweden, June 12th - 16th.


Pigeon, G., M.A. Moscicki, J.A. Voogt and V. Masson, 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.


Previous MEGAPOLI reports

Previous reports from the FP7 EC MEGAPOLI Project can be found at:
http://www.megapoli.info/

http://megapoli.dmi.dk/publ/MEGAPOLI_sr09-01.pdf

http://megapoli.dmi.dk/publ/MEGAPOLI_sr09-02.pdf

http://megapoli.dmi.dk/publ/MEGAPOLI_sr09-03.pdf

http://megapoli.dmi.dk/publ/MEGAPOLI_sr10-01.pdf

http://megapoli.dmi.dk/publ/MEGAPOLI_sr10-02.pdf

http://megapoli.dmi.dk/publ/MEGAPOLI_sr10-03.pdf

http://megapoli.dmi.dk/publ/MEGAPOLI_sr10-04.pdf

http://megapoli.dmi.dk/publ/MEGAPOLI_sr10-05.pdf

http://megapoli.dmi.dk/publ/MEGAPOLI_sr10-06.pdf
http://megapoli.dmi.dk/publ/MEGAPOLI_sr10-07.pdf
MEGAPOLI Project Partners

- DMI - Danish Meteorological Institute (Denmark) - Contact Persons: Prof. Alexander Baklanov (coordinator), Dr. Alexander Mahura (manager)
- FORTH - Foundation for Research and Technology, Hellas and University of Patras (Greece) - Prof. Spyros Pandis (vice-coordinator)
- MPIC - Max Planck Institute for Chemistry (Germany) - Dr. Mark Lawrence (vice-coordinator)
- ARIANET Consulting (Italy) – Dr. Sandro Finardi
- AUTH - Aristotle University Thessaloniki (Greece) - Prof. Nicolas Moussiopoulos
- CNRS - Centre National de Recherche Scientifique (incl. LISA, LaMP, LSCE, GAME, LGGE) (France) – Dr. Matthias Beekmann
- FMI - Finnish Meteorological Institute (Finland) – Prof. Jaakko Kukkonen
- JRC - Joint Research Center (Italy) – Dr. Stefano Galmarini
- ICTP - International Centre for Theoretical Physics (Italy) - Prof. Filippo Giorgi
- KCL - King’s College London (UK) – Prof. Sue Grimmond
- NILU - Norwegian Institute for Air Research (Norway) – Dr. Andreas Stohl
- PSI - Paul Scherrer Institute (Switzerland) – Prof. Urs Baltensperger
- TNO-Built Environment and Geosciences (The Netherlands) – Prof. Peter Buijtjes
- MetO - UK MetOffice (UK) – Dr. Bill Collins
- UHam - University of Hamburg (Germany) – Prof. Heinke Schluenzen
- UHel - University of Helsinki (Finland) – Prof. Markku Kulmala
- U-HCAIR - University of Hertfordshire, Centre for Atmospheric and Instrumentation Research (UK) – Prof. Ranjeet Sokhi
- USTUTT - University of Stuttgart (Germany) – Prof. Rainer Friedrich
- WMO - World Meteorological Organization (Switzerland) – Dr. Liisa Jalkanen
- UCam - Centre for Atmospheric Science, University of Cambridge (UK) – Prof. John Pyle

MEGAPOLI Project web-site
http://www.megapoli.info

MEGAPOLI Project Office
Danish Meteorological Institute (DMI)
Lyngbyvej 100, DK-2100
Copenhagen, Denmark
E-mail: alb@dmi.dk
Phone: +45-3915-7441
Fax: +45-3915-7400

Work Packages
WP1: Emissions
(H. Denier van der Gon, P. Buijtjes)
WP2: Megacity features
(S. Grimmond, I. Esau)
WP3: Megacity plume case study
(M. Beekmann, U. Baltensperger)
WP4: Megacity air quality
(N. Moussiopoulos)
WP5: Regional and global atmospheric composition
(J. Kukkonen, A. Stohl)
WP6: Regional and global climate impacts
(W. Collins, F. Giorgi)
WP7: Integrated tools and implementation
(R. Sokhi, H. Schlunzen)
WP8: Mitigation, policy options and impact assessment
(R. Friedrich, D. van den Hout)
WP9: Dissemination and Coordination
(A. Baklanov, M. Lawrence, S. Pandis)