Key conclusions of the first international urban land surface model comparison project

Article

Accepted Version

Best, M. J. and Grimmond, C. S. B. (2015) Key conclusions of the first international urban land surface model comparison project. Bulletin of the American Meteorological Society, 96 (5). pp. 805-819. ISSN 1520-0477 doi: https://doi.org/10.1175/BAMS-D-14-00122.1 Available at http://centaur.reading.ac.uk/37718/

Publisher: American Meteorological Society

Key conclusions of the first international urban land surface model comparison project

M.J. Best, C.S.B. Grimmond

AFFILIATIONS: Best – Met Office, Exeter, UK, and King's College London, London, UK; Grimmond – University of Reading, Reading, UK

CORRESPONDING AUTHOR: Martin J. Best, Met Office, FitzRoy Road, Exeter, EX1 3PB, UK.

E-mail: martin.best@metoffice.gov.uk

ABSTRACT: The first international urban land surface model comparison was designed to identify three aspects of the urban surface-atmosphere interactions: (1) the dominant physical processes, (2) the level of complexity required to model these, and (3) the parameter requirements for such a model. Offline simulations from 32 land surface schemes, with varying complexity, contributed to the comparison. Model results were analysed within a framework of physical classifications and over four stages. The results show that the following are important urban processes; (i) multiple reflections of shortwave radiation within street canyons, (ii) reduction in the amount of visible sky from within the canyon, which impacts on the net long-wave radiation, (iii) the contrast in surface temperatures between building roofs and street canyons, and (iv) evaporation from vegetation. Models that use an appropriate bulk albedo based on multiple solar reflections, represent building roof surfaces separately from street canyons and include a representation of vegetation demonstrate more skill, but require parameter information on the albedo, height of the buildings relative to the width of the streets (height to width ratio), the fraction of building roofs compared to street canyons from a plan view (plan area fraction) and the fraction of the surface that is vegetated. These results, whilst based on a single site and less than 18 months of data, have implications for the future design of urban land surface models, the data that need to be measured in urban observational campaigns, and what needs to be included in initiatives for regional and global parameter databases.

Capsule Summary

The conclusions from the first international urban land surface model comparison project have implications for future models, observations and parameter databases, that extend beyond the urban modelling community

1. Introduction

Urban areas are often warmer than their surrounding rural environments, referred to as the urban heat island (UHI). This urban warming has numerous effects, including the initiation of convective storms (e.g., *Bornstein and Lin*, 2000), altering pollution dispersion by adapting mixing through changes to atmospheric boundary layer structure (e.g., *Sarrat et al.*, 2006, *Luhar et al.*, 2014), impacts on the production and mixing of ozone (e.g., *Chaxel and Chollet*, 2009, *Ryu et al.*, 2013), enhanced energy demand for summer-time cooling through air conditioning (e.g., *Radhi and Sharples,* 2013, *Li et al.*, 2014), impacts on urban ecology (e.g., *Pickett et al.*, 2008, *Francis and Chadwick*, 2013) and increased mortality rates during heat waves (e.g., *Laaidi et al.*, 2012, *Herbst et al.*, 2014, *Saha et al.*, 2014). As such, it is important to be able to accurately forecast urban warming and other meteorological variables for cities where the majority of the World’s population now lives.

Predictions of future climate suggest additional warming in urban environments (*McCarthy et al.*, 2010, *Oleson et al.*, 2011). Indeed, the Inter-Governmental Panel on Climate Change (IPCC) Working Group 1 Fifth Assessment Report (*IPCC*, 2013) included at least one model that explicitly included an urban representation, and this number is likely to increase in the future as the resolution of these climate models increases to the extent that some urban areas are resolved. For future design of buildings and planning of cities, it is important that the dominant processes that lead to urban warming effects are considered. This requires the development of models that can represent the most important features of the urban heat island be used for reliable predictions.

The urban heat island results from differences in surface energy exchanges between the urban environment and its surrounding rural area. Thus, understanding these differences is needed to interpret the urban heat island. The differences in urban surface energy exchanges arise through a number of processes. The geometry of a street canyon will increase the incoming solar radiation and long-wave radiation that are absorbed, due to multiple reflections and re-radiated from the 3-dimensional structures. The orientation of street canyons and the elevation of the sun will impact the reflected solar radiation, as a consequence of the depth to which the direct sunshine can penetrate into the canyon. The reduced availability of water at the urban surface, compared to natural vegetated or bare soil surfaces, means more of the incoming solar radiation is transformed into heat rather than a flux of moisture into the atmosphere. However, a larger proportion of this energy for heating is held within the fabric of the buildings given the large thermal inertia of the materials, resulting in changes in the diurnal cycle of urban temperatures. Moreover, an additional source of heating within the urban areas comes from human activities such as transport, the internal heating of the buildings and the metabolic rates of the people themselves (e.g., Sailor and Lu, 2004).

All of these processes contribute to the differences in the energy balance between urban and rural surfaces, but it is difficult to identify which are the dominant processes just from observations as the processes cannot be separated because of the complex nature of the environment. As such, the best way to study these processes individually is by using urban land surface models (ULSMs) that have been developed for weather and climate applications, i.e., exchange surface fluxes with an atmospheric model. There are a number of such ULSMs that vary considerably in their complexity (e.g., *Kusaka et al.*, 2001, *Fortuniak*, 2003, *Krayenhoff and Voogt*, 2007, *Hamdi and Masson*, 2008, *Lee and Park*, 2008, *Oleson et al.*, 2008a). Although newer models often include more complex features than previous models, without knowing the dominant processes and controls, it is difficult to quantify the impact of each new feature.

The first urban land surface model comparison was designed to objectively assess and compare the performance of a range of ULSMs for a single observational site. It attempted to identify the dominant physical processes that need to be represented in ULSMs by comparing models of varying complexity (Table 1). These models ranged from simple bulk representations of the surface that have been applied to atmospheric models for over a decade, representations of the facets of a street canyon (i.e., roofs, walls and road) that have been used in weather and climate models, through to more recently developed schemes that consider a complete energy balance at various levels within the urban canyon that have been applied to stand alone single point studies. Figure 1 shows a conceptual representation of the surface energy balance for these models of varying complexity. Whilst the scale that these models typically represent is larger than the size of the elements within a street canyon, a common feature is the ability to predict the exchange of fluxes between the urban surface and the atmosphere above it, i.e., the net all-wave radiation (Q^*), turbulent sensible (Q_H) and latent heat

(Q_E) fluxes, as measured from flux towers in numerous urban observational campaigns.

The aim of the urban model comparison was to consider:

(1) What are the dominant physical processes in the urban environment?

(2) What is the level of complexity required for an ULSM to be fit for purpose?

(3) What are the parameter requirements for such a model?

Here we present an analysis of the model comparison results to address these questions.

2 Model Comparison design

2. 1 Observational data

The criteria for selecting the evaluation dataset were; first it had not been used to evaluate any ULSMs previously, and second it needed to cover an annual cycle to allow assessment for different seasons. Model evaluation studies often result in the development and optimisation of a model in order to obtain better representation of the assessed metrics. Hence, using a dataset previously used by one or a sub-set of the models to be evaluated would not enable a clean/independent objective assessment for all of the models.

The dataset for a suburb of Melbourne (Preston) (*Coutts et al.*, 2007a, 2007b) that had observations from 13 August 2003 to 13 November 2004 was selected. The moderately developed, low-density housing area is classified by *Coutts et al.* (2007b) as an Urban Climate Zone (UCZ) 5 (*Oke*, 2006), Local Climate Zone (LCZ) 6 (*Stewart and Oke*, 2012) or *Loridan and Grimmond* (2012) Urban Zone for Energy exchange (UZE) medium density. The description of UCZ 5 is "medium development, low density suburban with 1 or 2 storey houses, e.g., suburban housing" (*Oke*, 2006), and as such the site is typical of suburban areas found in North America, Europe and Australasia. The area has mean building height-to-width ratio of 0.42 and mean wall-to-plan ratio of 0.4 (*Coutts et al.*, 2007b). The surface is dominated by impervious cover (44.5% buildings, 4.5% concrete and 13% roads), with a pervious cover of 38% (15% grass, 22.5% other vegetation and 0.5% bare ground or pools) (*Coutts et al.*, 2007a).

The methods used to obtain the observed fluxes applied to our current analysis are given in Table 2, with details (e.g., data processing) presented in the original observation papers (*Coutts et al.*, 2007a, 2007b). In addition, the initial model comparison results papers (*Grimmond et al.*, 2011, *Best and Grimmond*, 2013, 2014) provide the site parameters. A continuous gap-filled atmospheric forcing dataset (474 days) to run the models was created for this study (see *Grimmond et al.*, 2011). To evaluate the modelled fluxes (sensible heat flux, latent heat flux, net all-wave radiative flux and net storage heat flux (ΔQ_S)) 30 min periods are used when no observed fluxes are missing to allow consistent analysis between the fluxes (N=8865 or 38.9% of the full period).

2.2 Data analysis

To permit the research questions posed above to be considered, information about the observational site was released to the modelling groups in stages. This enabled analysis of the importance of the different types of information to model performance through assessment of the change in model skill between the stages. The stages (Table 3), designed to correlate with ease of access to information for all cities globally, involved release of (*Grimmond et al.*, 2011):

Stage 1: Atmospheric forcing data: (Table 3), typically provided by an atmospheric model.

Stage 2: Vegetation and built fraction: two dimensional plan area characteristics of the site. These can be determined from land cover datasets derived from satellite data.

Stage 3: Morphology: three dimensional characteristics of the site (Table 3.). These can be interpreted from LiDAR (e.g., *Goodwin et al.*, 2009, *Lindberg and Grimmond,* 2011), aerial photographs (e.g., *Ellefsen*, 1990/1991), detailed satellite imagery (e.g., *Brunner et al.*, 2010), or simple empirical relations (e.g., *Bohnenstengel et al.*, 2011).

Stage 4: Building material parameters (Table 3): only obtainable from local knowledge of the materials used in the construction of the buildings.

Stage 5: Observed fluxes: to allow parameter optimisation studies. Only a few groups completed this stage, so these results are not presented here.

The results from 24 modelling groups are analysed, involving 21 independent models (Table 1). Alternative versions of the same model were run by the same or independent modelling groups, which resulted in 32 sets of model simulations being submitted for all of the four stages (see full list in *Grimmond et al.*, 2011). Each group completed a survey indicating the level of complexity used for various physical processes within their models. From the latter, categories of physical processes were established, with classes that cover the range of complexities (*Grimmond et al.*, 2010, 2011). These categories were chosen to investigate the importance of various physical processes that could contribute to differences in the surface energy balance between the urban and rural environments. Thus every model is assigned to a class in each category based on the survey information. In this study, the complexity category

(Grimmond et al., 2011) is not considered as the focus is to separate the specific physical processes. The categories, with the number of models in each class are shown in Table 4.

Comparing the mean behaviour of the models in each of the classes as a reference provides a method to determine the level of complexity that gives the best performance for each category. These data are analysed to address the second research question, where “fit for purpose” in this study is defined as being able to accurately represent the energy exchange between the urban surface and the atmosphere (i.e., the net all-wave radiation, turbulent sensible and latent heat fluxes).

Furthermore, by assessing the performance of the models across the categories for all classes, it is possible to identify the physical processes that have the largest impact on the performance of the models, hence identifying the dominant physical processes and addressing the first research question.

2.3 Methodology

Initial results from the urban model comparison (*Grimmond et al.*, 2011) ranked the models and assessed the performance of the various classes within the categories using standard statistical measures. Here an alternative approach to assess the models’ performance is used, that considers the percentage of the models’ data values that are within observational error (E_{obs}). This gives a measure between zero (no values within observational errors) and 100% (all values within observational errors, i.e., a ‘perfect’ model). Although this type of analysis is not strictly benchmarking, as each model is not being compared to an *a priori* metric, it could be considered as being closer to the benchmarking ethos as having all data points within observational errors would be a stringent metric.

The observational error estimates used in this analysis are for day-time fluxes based on a percentage of the observed fluxes, as suggested by *Hollinger and Richardson* (2005): net all-wave radiation flux 5%, turbulent sensible heat flux 10%, latent heat flux 8%, and upward components of both shortwave and long-wave radiation fluxes 10%. As the net storage heat flux in the observational dataset is determined as the residual of the surface energy balance, its observational error is assumed to be the sum of the errors for the other terms (i.e., Q^*, Q_H and Q_E), giving 23%. The night-time error estimates are assumed to be double the day-time error estimates for each of the fluxes. The absolute magnitude of fluxes during this period are typically small (order of (10) W m^{-2}), hence changes in the percentage of the observed flux used as the error estimates are likely to be within the reporting resolution (e.g. order of (1) W m^{-2}) of the observations (especially the turbulent fluxes). Whilst these error estimates may be indicative rather than the actual values, the results would not substantially change the analysis presented.

The analysis was undertaken for each model (*k*) in each class (*j*) within each category (*i*) (Table 4), for each flux, over each stage within the comparison, and separately for day-time and night-time. From this the percentage of data within observational error ($E_{obs\ i,j}$) was determined:

$$E_{obs,ij} = \frac{\sum_{k=1}^{n_{ij}} M_k}{n_{ij} T} \times 100\% \qquad (1)$$

where M is the number of points within observational error for model (*k*), *n* is the number of models and *T* is the number of day-time or night-time points in the time series as appropriate.

3. Results

Application of eqn. 1 to the sensible, latent and net storage heat fluxes, for each class and category, at Stage 1 and Stage 4 (Table 3) are shown in Figure 2. The results could range between 0% (i.e., no model data points within the observations errors) to 100% (i.e., all model data points within observational errors). The relative changes between the stages are also shown if Figure 2, i.e., for stage (s) the change relative to the previous stage (s-1) given by:

$$E^{s}_{obs,ij} / E^{s-1}_{obs,ij} \qquad (2)$$

Assessment of “between stages performance” allows an emphasis of the common results across all of the classes and categories. It is scaled between 0% and 100%, with 50% corresponding to no change between the stages (Figure 2).

Generally the results of the analysis, consistent with *Grimmond et al.* (2011), show that the skill to model latent heat fluxes is improved between stages 1 and 2. Knowing the plan area vegetation fraction (provided in Stage 2) is important for modelling the latent heat flux. No other stages show a general increase in model performance across the classes and categories for the fluxes shown in Figure 2. For the radiation fluxes (Fig. 3), the largest changes evident between Stages 3 and 4 are for the reflected shortwave radiation flux and are due to the specification of the bulk albedo at the site (i.e., the ratio of the reflected outgoing shortwave radiation flux from the whole urban surface to the incoming shortwave radiation flux, information released at Stage 4). This is also consistent with the conclusions from *Grimmond et al.* (2011).

Model performance for the outgoing long-wave radiation flux has its largest changes at night-time between Stages 3 and 4 (when the 3-d site morphological information (Table 3) were made available, Fig. 3). This enhanced performance at night could be related to improved estimates of the sky view factor which influences radiative trapping, and/or from improved estimates of the difference in nocturnal surface temperatures between building roofs and those of the roads and walls of the urban canyons. Improved performance is not detected in the day-time outgoing long-wave radiation flux (Fig. 3), probably because of the dominance of shortwave radiation at this time. These results were not identified in *Grimmond et al.* (2011) as there was no separate analysis for day-time and night-time.

It is evident from Figures 2 and 3 that the performance of the models for each of the fluxes does not improve consistently for each stage, as might be expected. This suggests that the models are not able to correctly make use of all of the information that is provided at each of the stages and hence the design of the models, and the use of their specific parameters, is not necessarily correct. This is discussed further in Grimmond et al. (2011).

Each model is assigned to one class for every category (Table 4). This means that a model with particularly good (or poor) performance will influence the results for its class in each of the categories. The implications of this are that it is not possible to ensure that the good performance from a particular class within one category is not actually resulting from the results of a class from a different category. This potential contamination of results by categories inhibits the analysis of the dominant physical processes and the suitability of the models. Both the analysis presented in *Grimmond et al.* (2011) and that in Figures 2 and 3 have this limitation, hence we will not consider further any results in Figures 2 and 3 for any specific class or category. Alternatively, to address this issue of cross-contamination, we repeat the complete analysis using eqn. 1 separately for each category (c), but only considering the subset of models from class (*a*). Hence for each class (*j*) in category (*i*) for the analysis of eqn. 1, the models used are those that are in both class (*a*) of category (*c*) and class (*j*) of category (*i*), of which there are $n_\alpha = n_{ca} \bigcap n_{ij}$, thus:

$$E_{obs,caij} = \frac{\sum_{k=1}^{n_\alpha} M_k}{n_\alpha T} \times 100\% \qquad (3)$$

This gives the equivalent of 26 versions of Figures 2 and 3 (one for each class in each category); although for a given subset of models it is inevitable that some classes will not have any members and hence have no data. We then apply the following equation for each of the stages to determine which of the original class of models has the best performance:

$$P_{ca} = \frac{\sum N_m}{N_{tot} - \left(\sum N_{nd}\right) - 1} \times 100\% \qquad (4)$$

where P_{ca} is the percentage of classes in the analysis that are improved from just the subset of models (compared to the analysis with the full set of models),

$$N_m = \sum_{k=1}^{N_{tot}} \begin{cases} 1 \text{ if } E_{obs,caij} > E_{obs,ij} \\ 0 \text{ otherwise} \end{cases} \qquad (5)$$

is the number of classes that are improved in the analysis, N_{tot} is the total number of classes ($\sum ij = 26$) and

$$N_{nd} = \sum_{k=1}^{N_{tot}} \begin{cases} 1 \text{ if } n_{ca} \bigcap n_{ij} = 0 \\ 0 \text{ otherwise} \end{cases} \qquad (6)$$

is the number of classes with no data.

Hence values of P_{ca} close to 100% relate to nearly all classes in all categories being improved from the physical process represented in class (a) of category (c). This indicates that this process and its representation are important to model performance. Whereas values close to 0% relate to almost all classes in all categories being degraded, suggesting that the representation of the physical process is detrimental to model performance. Values around 50% have a similar number of classes that are improved and degraded, suggesting that the representation of the physical process has little impact on model performance. Hence the conclusions that can be drawn from this analysis are more robust than those of Figures 2 and 3, and the previous study of *Grimmond et al.* (2011).

For example, with models that have an infinite number of reflections (category R, class i), the median of the results over the stages give a value of 88% for the night-time net storage heat flux (Fig. 4). This results from 14 of the 16 possible classes containing data being improved when considering only these models, demonstrating that this is important for predicting this flux. However, models that have multiple reflections (category R, class m) have a value of 12.5% for the night-time net storage heat flux (Fig. 4). This results from only two of the possible 16 classes containing data being improved, hence showing that this is detrimental to predicting the flux.

The results of Figure 4 show that for some classes (e.g., infinite reflections; category R, class I, Table 4), there are some demonstrated improvements to a flux (e.g., LW_{up}) which is not obviously explained by the physics (e.g., how do infinite reflections of shortwave radiation improve the outgoing long-wave radiation but not the reflected shortwave?). Also, there are some classes that improve one particular flux, but not other fluxes. For example, models that represent the net storage heat flux as the residual of the surface energy balance (category S, class r, Table 4) demonstrate a clear improvement for the day-time sensible heat flux, but not for the latent or the net storage heat fluxes. This could be because with such models the sensible heat flux is not constrained by the energy balance giving them the freedom to enable better predictions of the sensible heat flux, whilst moisture availability is still the main control for the latent heat flux.

There are many such conclusions that can be drawn from Figure 4. Here the focus is on results that are consistent between the fluxes, or consistent for a particular flux between the day-time and night-time.

Models with a bulk representation of the albedo and emissivity (category A_E, class 1, Table 4), and a bulk representation of facets and orientation (category F_O, class 1; the models in these two classes were identical), demonstrate an improvement in skill during the day-time for nearly all fluxes, with the exceptions of the outgoing long-wave radiation which shows little change in skill and net all-wave radiation fluxes with only small improvements (Fig. 4). This class of models also shows an improvement in the night-time sensible and latent heat fluxes, but degradation in the radiative fluxes during the night. These improved results are most likely due to the ability to utilize the observed bulk albedo directly. This class of models clearly delivers the largest benefits across the fluxes and indicates the most significant physical process to represent is the bulk albedo for the urban surface, because the net shortwave radiation dominates the surface energy balance.

Improvements to the outgoing long-wave radiation flux and the net all-wave radiation flux during both day-time and night-time are obtained from models that have a single layer for each element of the urban environment (i.e., roofs and either urban canyons, or walls and roads separately) in the morphology category (category L, class 2, Table 4; Fig. 4). Improvements to the night-time sensible heat flux and net storage heat flux are also obtained from this class of models, but there is no improvement to these fluxes during the day-time. This neutral day-time result in the sensible and net storage heat fluxes may be explained by the negative impact on the outgoing shortwave radiation flux, which dominates over the long-wave radiation flux during the day-time. However, these results demonstrate the importance of presenting the difference in radiative surface temperatures between the roofs and the urban canyon, due to the non-linear relationship between the upward long-wave radiation and the radiative temperature.

When considering the way in which the models represent vegetation (category V, Table 4), we find that although including vegetation (classes s and i, Table 4) does generally lead to an improvement for the fluxes, these improvements are not as obvious as those from the bulk albedo or the single layer urban morphology. Hence although these results confirm those presented in earlier studies on the comparison (*Grimmond et al.*, 2011, *Best and Grimmond*, 2013, 2014), that representing vegetation gives improved results, we demonstrate that the more robust analysis presented here shows that this is not the most important physical process as was concluded in these earlier studies. Getting the radiative fluxes correct from the shortwave via the bulk albedo and the long-wave through the urban morphology are required before the vegetation can influence the partitioning of energy between the sensible and latent heat fluxes.

Previous studies on the urban comparison data have also concluded that models which neglect the anthropogenic heat flux (Q_F) do at least as well as the models that include this flux, although they were unable to explain this result (*Grimmond et al.*, 2011, *Best and Grimmond*, 2013, 2014). However, the results in Figure 4 show that although the class of models that neglect the anthropogenic heat flux (category A_N, class n, Table 4) do improve some of the fluxes, the improvements are not consistent over all of the fluxes. Moreover, this class of models within the anthropogenic heat flux category is not always the one that delivers the best results. Hence we can conclude that although the models that neglect the anthropogenic heat flux do show some improved results, we cannot make any significant statements about the classes within this category.

4. **Conclusions**

Prior conclusions from the ULSM comparison with daily (24 h) and seasonal analysis include that: representation of vegetation is critical to model performance (*Grimmond et al.,* 2011, *Best and Grimmond*, 2013), along with the associated initial soil moisture (*Best and Grimmond*, 2014), and the bulk albedo is also important (Grimmond et al., 2011). Notably, neglecting the distinctive urban anthropogenic heat flux was not found to penalize performance (albeit in the suburban area the value is small) (*Best and Grimmond*, 2013). However, this new analysis considering diurnal performance (day, night) enables us to conclude that nocturnal radiative processes also benefit from accounting for the enhanced long-wave trapping that occurs within urban areas. Separating the radiative processes of the roof and the urban canyon is beneficial.

More critically, the more robust analysis presented here enables identification of a re-prioritisation of the key physical processes: firstly, ensuring the use of the correct bulk albedo for the urban surface; secondly, the outgoing long-wave radiative fluxes with the representation of morphology separated into roofs and urban canyons; and thirdly, the inclusion of vegetation. The implications of the bulk albedo is important for observations as the temporal resolution of satellite estimates mean they will not provide the variations by time of day that are observed (e.g., Christen and Voogt, 2004, Grimmond et al. 2004, Kotthaus and Grimmond 2014).

The current results for anthropogenic heat flux are consistent with the earlier studies: that neglect of the relatively small magnitude flux at this site (study period mean = ~17 W m^{-2}) is reasonable. This conclusion could well be different for urban environments where this is a more significant term in the surface energy balance. The flux is expected to be larger in other areas of Melbourne (e.g., as suggested from analysis using the model of Lindberg et al. 2013) and for urban areas elsewhere. We therefore recommend that future model comparisons ideally include areas of cities with larger anthropogenic heat fluxes.

Thus to answer the three over-arching research questions of the urban model comparison:

(i) The dominant physical processes in the urban environment that models need to be able to simulate, in order, are; changes to the bulk albedo of the surface that result from building materials and also shortwave trapping from the canyon geometry; the reduction in outgoing long-wave radiation from the street canyon due to a reduced sky view factor and the contrast between this and the roofs that see a full sky view; and the evaporation from vegetation.

(ii) For the current generation of ULSMs, the ability to utilize a bulk surface albedo (category A_E, class 1, Table 4) and to be able to distinguish between the roofs of buildings and the urban canyons (category L, class 2), and to have a representation of vegetation (category V, classes s, i), results in the best performance.

(iii) The key parameters for ULSMs are the bulk surface albedo (information given for Stage 4 influencing the upward shortwave radiation flux), the height to width ratio of the urban canyons and the fraction of building roofs to the urban canyons (information given for Stage 3 influencing the upward long-wave radiation flux), and the vegetation fraction (information given for Stage 2 influencing the sensible and latent heat fluxes).

The results, from this and the previous studies on the ULSM comparison, all suggest that a simple representation for most of the physical categories is sufficient for this type of application, i.e., determination of local scale fluxes (e.g. for use in the coupling to an atmospheric model). The prior categorization of the models (Grimmond et al., 2011, Best and Grimmond, 2013) into (simple, medium and complex) complexity classes based upon the number of physical categories treated as simple by a model demonstrated that the simple models performed best. This relative success of simple models suggests that for simulating local scale fluxes, more complex schemes deliver little additional benefit. Furthermore, the reduced parameter requirements for simple schemes are advantageous for large scale applications, such as global or regional scale modelling. However, it cannot be expected that this conclusion would also hold for other applications, e.g., atmospheric dispersion within street canyons of a specific city, as the simple models do not present some of the basic physical requirements for such applications. Thus the requirement for the development of more complex ULSMs does remain.

The implications of this study go beyond the urban environment. In general, we need to balance the requirement for complexity within models against what is actually required for a model to be fit for purpose. Hence new and more complex processes should not be included in models unless it can be demonstrated that they are required. In addition, consideration needs to be given to the availability of information to specify parameters within complex models, and if such complexity can be justified given the uncertainty range for the parameters. Also, the type of analysis used here could be applied to any comparison study to ensure that the results are robust and not contaminated by physical processes not being directly considered.

These key conclusions are based on the single site observational dataset of less than 18 months. This suburban site of low density housing, is typical of extensive areas in North America, Europe and Australasia. Hence we might expect the results from this study to be valid over a reasonable range of cities. However, most urban environments have a range of zones (e.g. *Ellefsen*, 1991, *Grimmond and Souch*, 1994, *Stewart and Oke*, 2012) with very different characteristics. So to test if the results presented here are robust for other cities, similar “experiments” are required for additional sites with differing climates and urban characteristics. Hence we recommend that further model comparison projects are required for the urban community.

Despite these limitations, the results have implications for future development of ULSMs and for the types of data that need to be collected in future urban measurement campaigns (e.g., soil moisture, given its impact to limit transpiration and the long timescales required for model spin-up, along with the conclusion that the fraction of vegetation is important for urban areas) and/or the parameters that should be collated systematically for cities around the world (e.g., *Ching et al.*, 2009, *Loridan and Grimmond*, 2012, *Stewart and Oke*, 2012, *Ching*, 2013, *Faroux et al.*, 2013).

Acknowledgements Funds to support the comparison project were provided by the Met Office (P001550). M. Best was supported by the Joint DECC/Defra Met Office Hadley Centre Climate Programme (CA01101). We would like to thank Andrew Coutts and Jason Beringer for allowing their data to be used for the comparison. We would also like to thank Mathew Blackett for all of his efforts in coordinating the model data collection, and to everyone who contributed results to the comparison from their models: J.-J. Baik, S.E. Belcher, J. Beringer, S.I. Bohnenstengel, I. Calmet, F. Chen, A. Dandou, K. Fortuniak, M.L. Gouvea, R. Hamdi, M. Hendry, M. Kanda, T. Kawai, Y. Kawamoto, H. Kondo, E.S. Krayenhoff, S.-H. Lee, T. Loridan, A. Martilli, V. Masson, S. Miao, K. Oleson, R. Ooka, G. Pigeon, A. Porson, Y.-H. Ryu, F. Salamanca, G.J. Steeneveld, M. Tombrou, J.A. Voogt, D.T. Young and N. Zhang.

References

Best, M.J. (2005), Representing urban areas within operational numerical weather prediction models. *Boundary-Layer Meteorol*, 114: 91–109.

Best M.J., C.S.B. Grimmond, M.G. Villani (2006), Evaluation of the urban tile in MOSES using surface energy balance observations. *Boundary- Layer Meteorol,* 118: 503–525.

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

Best, M.J., C.S.B. Grimmond (2013), Analysis of the seasonal cycle within the first international urban land surface model comparison, *Boundary-Layer Meteorol.,* 146, 421-446, doi: 10.1007/s10546-012-9769-7.

Best, M.J., C.S.B. Grimmond (2014), Importance of initial state and atmospheric conditions for urban land surface models performance, *Urban Climate*. In press, doi:10.1016/j.uclim.2013.10.006

Bohnenstengel, S.I., S. Evans, P.A. Clark, S.E. Belcher (2011), Simulations of the London urban heat island, *Q. J. R. Meteorol. Soc.*, 137, 1625-1640, doi: 10.1002/qj.855.

Bornstein, R., Q. Lin (2000), Urban heat islands and summertime convection thunderstorms in Atlanta: three case studies, *Atmos. Environ.*, 34, 507-516, doi:10.1016/S1352-2310(99)00374-X.

Brunner, D., G. Lemoire, L. Bruzzone, H. Greidonus (2010), Building height retrieval from VHR SAR imagery based on an iterative simulation and matching technique. *IEEE Transactions on Geoscience and Remote Sensing*, 48, No.3, doi:10.1109/TGRS.2009.2031910.

Chaxel, E., J.-P. Chollet (2009), Ozone production from Grenoble city during the August 2003 heat wave, *Atmos. Environ.*, 43, 4784-4792, doi:10.1016/j.atmosenv.2008.10.054.

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

Ching, J.K.S. (2013), A perspective on urban canopy layer modelling for weather, climate and air quality applications, *Urban Climate*, 3, 13-39.

Ching, J., M. Brown, S. Burian, F. Chen, R. Cionco, A. Hanna, T. Hultgren, T. McPherson, D. Sailor, H. Taha, D. Williams (2009), National urban database and access portal tool, *Bull. American Meteorol. Soc.*, 90, 1157-1168, doi:10.1175/2009BAMS2675.1.

Christen, A., R. Voogt (2004), Energy and radiation balance of a central European city, Int. J. Climatol., 24, 1395-1421, doi:10.1002/joc.1074.

Coutts, A.M., J. Beringer, N.J. Tapper (2007a), Characteristics influencing the variability of urban CO2 fluxes in Melbourne, Australia, *Atmos. Environ.*, 41, 51-62.

Coutts, A.M., J. Beringer, N.J. Tapper (2007b) Impact of increasing urban density on local climate: spatial and temporal variations in the surface energy balance in Melbourne, Australia, *J. Appl. Meteorol.*, 47, 477-493.

Dandou A., M. Tombrou, E. Akylas, N. Soulakellis, E. Bossioli (2005), Development and evaluation of an urban parameterization scheme in the Penn State/NCAR Mesoscale model (MM5). *J Geophys Res*, 110: D10102. doi:10.1029/2004JD005192.

Dupont S., P.G. Mestayer (2006), Parameterisation of the urban energy budget with the submesoscale soil model. *J Appl Meteorol Climatol*, 45: 1744–1765.

Dupont S., P.G. Mestayer, E. Guilloteau, E. Berthier, H. Andrieu (2006), Parameterisation of the urban water budget with the submesoscale soil model. *J Appl Meteorol Climatol*, 45: 624–648.

Ellefsen, R. (1991), Mapping and measuring buildings in the canopy boundary layer in ten U.S. cities. Energy and Buildings, 16, 1025-1049.

Essery R.L.H., M.J. Best, R.A. Betts, P.M. Cox, C.M. Taylor (2003), Explicit representation of subgrid heterogeneity in a GCM land surface scheme. *J Hydrometeorol*, 4: 530–543.

Faroux, S., A. T. Kaptue Tchuente, J.-L. Roujean, V. Masson, E. Martin, P. Le Moigne (2013), ECOCLIMAP-II/Europe: a twofold database of ecosystems and surface parameters at 1 km resolution based on satellite information for use in land surface, meteorological and climate models, *Geosci. Model Dev.*, 6, 563-582, doi:10.5194/gmd-6-563-2013.

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

Fortuniak K., B. Offerle, 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*; 82–83. Available from: www.met.rdg.ac.uk/urb met/NATO ASI/talks.html (Last accessed 4–15 May 2010).

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

Francis, R.A., M.A. Chadwick. (2013), Urban Ecosystems: Understanding the Human Environment. *Routledge*, 220pp.

Goodwin, N.R., N.C. Coops, T.R. Tooke, A. Christen, J.A. Voogt (2009), Characterizing urban surface cover and structure with airborne lidar technology. *Can. J. Remote Sens.*, 35, 297-309.

Grimmond, C.S.B., C. Souch (1994), Surface description for urban climate studies: a GIS based methodology. *Geocarto. International*, 9, 47-59.

Grimmond C.S.B., T.R. Oke (2002), Turbulent heat fluxes in urban areas: observations and local-scale urban meteorological parameterization scheme (LUMPS). *J Appl Meteorol,* 41: 792–810.

Grimmond, C.S.B., J.A. Salmond, T.R. Oke, B. Offerle, A. Lemonsu (2004), Flux and turbulence measurements at a densely built-up site in Marseille: Heat, mass (water and carbon dioxide), and momentum. *J. Geophys. Res. Atmos*., 109, D24101, doi:10.1029/2004jd004936

Grimmond, C.S.B., M. Blackett, M.J. Best, J. Barlow, J.-J. Baik, S.E. Belcher, S.I. Bohnenstengel, I. Calmet, F. Chen, A. Dandou, K. Fortuniak, M.L. Gouvea , R. Hamdi, M. Hendry, T. Kawai, Y. Kawamoto, H. Kondo, E.S. Krayenhoff, S.-H. Lee, T. Loridan, A. Martilli, V. Masson, S. Miao, K. Oleson, G. Pigeon, A.

Porson, Y.-H. Ryu, F. Salamanca, L. Shashua-Bar, G.-J. Steeneveld, M. Trombou, J. Voogt, D. Young, N. Zhang (2010), The international urban energy balance models comparison project: first results from phase 1, *J. Appl. Meteorol. Climatol.*, 49, 1268-1292, doi: 10.1175/2010JAMC2354.1.

Grimmond, C.S.B., M. Blackett, M.J. Best, J.-J. Baik, S.E. Belcher, J. Beringer, S.I. Bohnenstengel, I. Calmet, F. Chen, A. Coutts, A. Dandou, K. Fortuniak, M.L. Gouvea , R. Hamdi, M. Hendry, M. Kanda, T. Kawai, Y. Kawamoto, H. Kondo, E.S. Krayenhoff, S.-H. Lee, T. Loridan, A. Martilli, V. Masson, S. Miao, K. Oleson, R. Ooka, G. Pigeon, A. Porson, Y.-H. Ryu, F. Salamanca, G.-J. Steeneveld, M. Trombou, J. Voogt, D. Young, N. Zhang (2011), Initial results from phase 2 of the international urban energy balance model comparison, *Int. J. Climatol.*, 30, 244-272, doi:10.1002/joc.2227.

Hamdi, R., V. Masson (2008), Inclusion of a drag approach in the Town Energy Balance (TEB) scheme: offline 1-D evaluation in a street canyon, *J. Appl. Meteorol. Climatol.*, 47, 2627-2644.

Harman I.N., M.J. Best, S.E. Belcher (2004a), Radiative exchange in an urban street canyon. *Boundary-Layer Meteorol*, 110: 301–316.

Harman I.N., J.F. Barlow, S.E. Belcher (2004b), Scalar fluxes from urban street canyons. Part II: model. *Boundary-Layer Meteorol*, 113: 387–410.

Harman I.N., S.E. Belcher (2006), The surface energy balance and boundary layer over urban street canyons. *Q J R Meteorol Soc*, 132: 2749–2768.

Herbst, J., K. Mason, R.W. Byard, J.D. Bilbert, C. Charlwood, K.J. Heath, C. Winskog, N.E.I. Langlois (2014), Heat-related deaths in Adelaide, South Australia: Review of the literature and case findings – An Australian perspective, *J. Forensic and Legal Medicine*, 22, 73-78. doi:10.1016/j.jflm.2013.12.018

Hollinger, D.Y., A.D. Richardson (2005), Uncertainty in eddy covariance measurements and its application to physiological models, *Tree Physiol.*, 25, 873-885.

IPCC (2014), Working Group I Contribution to the IPCC Fifth Assessment Report (AR5), Climate Change 2013: The Physical Science Basis. *Intergovernmental Panel on Climate Change*, Geneva, Switzerland.

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

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

Kawai T., M. Kanda, K. Narita, A. Hagishima (2007), Validation of a numerical model for urban energy-exchange using outdoor scalemodel measurements. *Int J Climatol*, 27: 1931–1942.

Kawai T., M.K. Ridwan, M. Kanda (2009), Evaluation of the simple urban energy balance model using 1-yr flux observations at two cities. *J Appl Meteorol Climatol*, 48: 693–715.

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

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

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

Kondo H., F.H. Liu (1998), A study on the urban thermal environment obtained through a one-dimensional urban canopy model, *J Jpn Soc Atmos Environ*. 33, 179-192 (in Japanese)

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

Kotthaus, S., C.S.B. Grimmond (2013), Energy exchange in a dense urban environment - Part II: Impact of spatial heterogeneity of the surface, Urban Climate, hppt://dx.doi/org/10.1016/j.uclim.2013.10.001.

Krayenhoff, E.S., J.A. Voogt (2007), A microscale three-dimensional urban energy balance model for studying surface temperatures, *Boundary-Layer Meteorol.*, 123, 433-461.

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

Laaidi, K., A. Zeghnoun, B. Dousset, P. Bretin, S. Vandentorren, E. Giraudet, P. Beaudeau (2012), The impact of heat islands on mortality in Paris during the August 2003 heat wave, *Environ. Health Perspectives*, 120, 254-259, doi:10.1289/ehp.1103532.

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

Lemonsu A., C.S.B. Grimmond, V. Masson (2004), Modelling the surface energy balance of an old Mediterranean city core. *J Appl Meteorol,* 43: 312–327.

Li, C.B., J.J. Zhou, Y.J. Cao, J. Zhong, Y. Liu, C.Q. Kang, Y. Tan (2014), Interaction between urban microclimate and electric air-conditioning energy consumption during high temperature season, *Applied Energy,* 117, 149-156, doi:10.1016/j.apenergy.2013.11.057.

Lindberg, F., C.S.B. Grimmond (2011), Nature of vegetation and building morphology characteristics across a city: Influence on shadow patterns and mean radiant temperatures in London. *Urban Ecosyst*., 14, 617-623. doi:10.1007/s11252-011-0184-5.

Lindberg, F., C.S.B. Grimmond, N. Yogeswaran, S. Kotthaus, L. Alen (2013), Impacts of city changes and weather on anthropogenic heat flux in Europe 1995-2015, *Urban Climate*, 4, 1-15. http://dx.doi.org/10.1016/j.uclim.2013.03.002.

Loridan T., C.S.B. Grimmond, S. Grossman-Clarke, F. Chen, M. Tewari, K. Manning, A. Martilli, H Kusaka, M. Best (2010), Trade-offs and responsiveness of the single-layer urban canopy parameterization in WRF: an offline evaluation using the MOSCEM optimization algorithm and field observations. *Q J R Meteorol Soc,* 136: 997–1019. doi:10.1002/qj.614.

Loridan T., C.S.B. Grimmond, B.D. Offerle, D.T. Young, T. Smith, L. Järvi, F. Lindberg (2011), Local-scale urban meteorological parameterization scheme (LUMPS): Longwave radiation parameterization and seasonality-related developments. *J Appl Meteorol Climatol,* 50: 185-202. doi:10.1175/2010JAMC2474.1

Loridan, T., C.S.B. Grimmond (2012), Multi-site evaluation of an urban land-surface model: intra-urban heterogeneity, seasonality, and parameter complexity requirements. *Q. J. R. Meteorol. Soc.*, 138, 1094-1113, doi:10.1002/qj.963.

Luhar, A.K., M. Thatcher, P.J. Hurley (2014) Evaluating a building-averaged urban surface scheme in an operational mesoscale model for flow and dispersion, *Atmos. Environ.*, 88, 47-58, doi:10.1016/j.atmosenv.2014.01.059.

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

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

Masson V., C.S.B. Grimmond, T.R. Oke (2002), Evaluation of the Town Energy Balance (TEB) scheme with direct measurements from dry districts in two cities. *J Appl Meteorol,* 41: 1011–1026.

McCarthy, M.P., M.J. Best, R.A. Betts (2010), Climate change in cities due to global warming and urban effects, *Geophys. Res. Letters*, 37, L09705, doi:10.1029/2010GL042845.

Offerle B., C.S.B. Grimmond, T.R. Oke (2003), Parameterization of net all-wave radiation for urban areas. *J Appl Meteorol,* 42: 1157–1173.

Oke TR (2006) Towards better scientific communication in urban climate. Theor. Appl. Climatol. 84: 179-190. doi: 10.1007/s00704-005-0153-0

Oleson, K.W., G.B. Bonan, J. Feddema, M. Vertenstein, C.S.B. Grimmond (2008a), An urban parameterization for a global climate model: 1. Formulation and evaluation for two cities, *J. Appl. Meteorol. Climatol.*, 47, 1038-1060.

Oleson K.W., G.B. Bonan, J. Feddema, M. Vertenstein. 2008b. An urban parameterization for a global climate model: 2. Sensitivity to input parameters and the simulated heat island in offline simulations. *J Appl Meteorol Climatol*, 47: 1061–1076.

Oleson, K.W., G.B. Bonan, J. Feddema, T. Jackson (2011), An examination of urban heat island characteristics in a global climate model, *Int. J. Climatol.*, 31, 1848-1865. doi:10.1002/joc.2201.

Pickett, S.T.A., M.L. Cadenasso, J.M. Grove, P.M. Groffman, L.E. Band, C.G. Boone, W.R. Burch, Jr., C.S.B. Grimmond, J. Hom, J.C. Jenkins, N.L. Law, C.H. Nilon, R.V. Pouyat, K. Szlavecz, P.S. Warren, M.A. Wilson. 2008. Beyond urban legends: an emerging framework of urban ecology, as illustrated by the Baltimore Ecosystem Study. *BioScience*. 58(2):139-150

Pigeon G., M.A. Moscicki, J.A. Voogt, V. Masson (2008), Simulation of fall and winter surface energy balance over a dense urban area using the TEB scheme. *Meteorol Atmos Phys*, 102: 159–171.

Porson A., P.A. Clark, I.N. Harman, M.J. Best, S.E. Belcher (2010), Implementation of a new urban energy budget scheme in the MetUM. Part II. Validation against observations and model intercomparison. *Q J R Meteorol Soc*, 136: 1530-1542.

Radhi, H., S. Sharples (2013), Quantifying the domestic electricity consumption for air-conditioning due to urban heat islands in hot arid regions, *Applied Energy*, 112, 371-380, doi:10.1016/j.apenergy.2013.06.013.

Ryu Y.-H., J.-J. Baik, S.-H. Lee (2011), A new single-layer urban canopy model for use in mesoscale atmospheric models. *J Appl Meteorol Climatol*, 50: 1773-1794. doi: 10.1175/2011JAMC2665.1

Ryu, Y.-H., J.J. Baik (2013), Effects of anthropogenic heat on ozone air quality in a megacity. *Atmos. Environ.*, 80, 20-30, doi:10.1016/j.atmosenv.2013.07.053.

Saha, M.V., R.E. Davis, D.M. Hondula (2014), Mortality displacement as a function of heat event strength in 7 US cities, *American J. Epidemiology*, 179, 467-474, doi:10.1093/aje/kwt264.

Sailor, D.J., L. Lu (2004), A top-down methodology for developing diurnal and seasonal anthropogenic heating profiles for urban areas. Atmos. Environ., 38, 2737-2748, doi:10.1016/j.atmosenv.2004.01.034.

Salamanca F., E.S. Krayenhoff, A. Martilli (2009), On the derivation of material thermal properties representative of heterogeneous urban neighbourhoods. *J Appl Meteorol Climatol*, 48: 1725–1732.

Salamanca F., A. Krpo, A. Martilli, A. Clappier (2010), A new building energy model coupled with an urban canopy parameterization for urban climate simulations – part I. Formulation, verification, and sensitivity analysis of the model. *Theor Appl Climatol*, 99: 345-356. doi: 10.1007/s00704-009-0142-9.

Salamanca F., A. Martilli (2010), A new Building Energy Model coupled with an Urban Canopy Parameterization for urban climate simulations – part II. Validation with one dimension off-line simulations. *Theor Appl Climatol*, 99: 345–356.

Sarrat, C., A. Lemonsu, V. Masson, D. Guedalla (2006), Impact of urban heat island on regional atmospheric pollution, *Atmos. Environ.*, 40, 1743-1758, doi:10.1016/j.atmosenv.2005.11.037.

Stewart, I.D., T.R. Oke (2012), Local climate zones for urban temperature studies, *Bull. American Meteorol. Soc.*, 93, 1879-1900, doi:10.1175/BAMS-D-11-00019.1.

Figure captions

Figure 1: Conceptual figure of how surface energy balance exchanges are included in urban land surface models of different complexity. Note individual models have simple and complex features (Grimmond et al., 2011).

Figure 2: For each flux and physical category class (Table 4), the percentage of modelled data points within the specified observational errors (eqn. 1) for Stages 1 and 4 (grey) plus the change relative to the previous stage (eqn. 2; scaled between -100% and 100%, shown by the horizontal dotted lines). Blue shading indicates an improvement (> 0) and red degradation (< 0). Results are shown for day and night-time (with day defined as incoming solar radiation flux greater than 0 W m^{-2}). Codes definition for the physical categories and component classes (used in the x-axis) are given in Table 4

Figure 3: As for Fig. 2, but for the radiative fluxes

Figure 4: The subset of models within a class of a category improved compared to all models (P_{ca}, eqn. 4) ranked according to the median over the stages (for each flux, by time of day (as for Fig. 2)). Shading shows the range of results over the stages, with the individual results shown as horizontal lines within this. The colouring emphasises the values of the median over the stages, with 100% corresponding to all classes improved, 0% all classes degraded and 50% no change. Note X-axis code (Table 4) order changes between subplots because of ranking (Colour text is to aid differences to be noted).

Table 1: Urban land surface models (ULSMs) used to obtain results that are analysed here. See Grimmond et al. (2010, 2011) for more details of the different model versions and the number of groups that submitted simulations to the urban model comparison.

Model name	References
Building effect parameterization (BEP)	Martilli et al. (2002) Salamanca et al. (2009, 2010) ; Salamanca and Martilli (2010)
Community Land Model – urban (CLM-urban)	Oleson et al. (2008a, 2008b)
Institute of Industrial Science urban canopy model	Kawamoto and Ooka (2006, 2009a, 2009b)
Joint UK land environment simulator (JULES)	Essery et al. (2003); Best (2005); Best et al. (2006); Best et al. (2011)
Local-scale urban meteorological parameterization scheme (LUMPS)	Grimmond and Oke (2002); Offerle et al. (2003); Loridan et al. (2011)
Met Office Reading urban surface exchange scheme (MORUSES)	Harman et al. (2004a, 2004b); Harman and Belcher (2006), Porson et al. (2010)
Multi-layer urban canopy model	Kondo and Liu (1998); Kondo et al. (2005)
National and Kapodistrian University of Athens model	Dandou et al. (2005)
Noah land surface model/single-layer urban canopy model	Kusaka et al. (2001); Chen et al. (2004); Loridan et al. (2010)
Seoul National University urban canopy model	Ryu et al. (2011)
Simple urban energy balance model for mesoscale simulation	Kanda et al. (2005a, 2005b); Kawai et al. (2007, 2009)
Slab urban energy balance model	Fortuniak (2003); Fortuniak et al. (2004, 2005)
Soil model for submesoscales (urbanized)	Duport and Mestayer (2006); Dupont et al. (2006)
Temperatures of urban facets (TUF)	Krayenhoff and Voogt (2007)
Town energy balance (TEB)	Masson (2000); Masson et al. (2002); Lemonsu et al. (2004); Pigeon et al. (2008), Hamdi and Masson (2008)
Vegetated urban canopy model	Lee and Park (2008)

Table 2: Methods used to obtain the observed fluxes used for comparison with the ULSM. Sources: *Coutts et al*., (2007a, 2007b). Height of observation for all fluxes: 40 m.

Flux	Instrument / Method	Sampling frequency (Hz.)	Averaging period (min)
SW_{up} LW_{up} Q^*	Kipp and Zonen CM7B and CG4 radiometers	1	30
Q_H	CSI CSAT 3D sonic anemometer	10	30
Q_E	CSI CSAT 3D sonic anemometer CSI Krypton hygrometer (Aug 2003 – Feb 2004), LiCOR LI7500 open-path infrared gas analyser (remaining period)	10	30
ΔQ_S	Residual of the surface energy balance	N/A	30
Q_F	Calculated (*Sailor and Lu* ,2004) : *Vehicles*: Numbers from survey (Nov. 2002 – Oct 2003 *Building* sector: 30 min electricity and daily natural gas statistics *Human metabolism*: Night, day and transition period metabolic rates, with population density statistics	N/A	Average monthly diurnal cycle at 30 min. resolution

Table 3. Information released at each stage of the comparison

Stage	Information released
1	Atmospheric forcing data only (incoming shortwave radiation, incoming long-wave radiation, precipitation, atmospheric wind speed, temperature, specific humidity and surface pressure)
2	Vegetation and built fractions
3	Morphology (Building heights, height-to-width ratio, mean wall to plan area ratio, fraction of surface covered by buildings, concrete, road,)
4	Specific information on building materials (e.g., albedo and thermal properties of wall, road, roof)
5	Observed fluxes for parameter optimisation (Not considered in this study)

Table 4: Classes and physical categories used in the analysis of the urban comparison results, including the number of models in each class (see *Grimmond et al.*, 2010, 2011 for more details). Colours are used on the plots to aid comparison.

Category	**Class**			
Vegetation (V)	None (**n**)	Separate tile (**s**)	Integrated (**i**)	
No. of models	8	19	5	
Anthropogenic heat flux (AN)	None (**n**)	Prescribed flux (**p**)	Internal building temperature (**i**)	Modelled (**m**)
No. of models	22	2	6	2
Temporal variation of the anthropogenic heat flux (T)	None (i.e., no flux) (**n**)	Fixed (i.e., time invariant flux) (**f**)	Variable (i.e., time varying flux) (**v**)	
No. of models	22	3	7	
Urban morphology (L)	Bulk (**1**)	Single layer (**2**)	Multiple layer (**4**)	
No. of models	6	20	6	
Facets & orientation (FO)	Bulk (**1**)	Roof, walls, road without orientation (**n**)	Roof, walls, road with orientation, no intersections (**o**)	Roof, walls, road with orientation and intersections (**i**)
No. of models	5	17	6	4
Reflections (R)	Single (**1**)	Multiple (**m**)	Infinite (**i**)	
No. of models	11	13	8	
Albedo, emissivity (AE)	Bulk (**1**)	Two facet (**2**)	Three facet (**3**)	
No. of models	5	4	23	
Net storage heat flux (S)	Net all wave radiation (**n**)	Surface energy balance residual (**r**)	Conduction equation (**c**)	
No. of models	3	6	23	

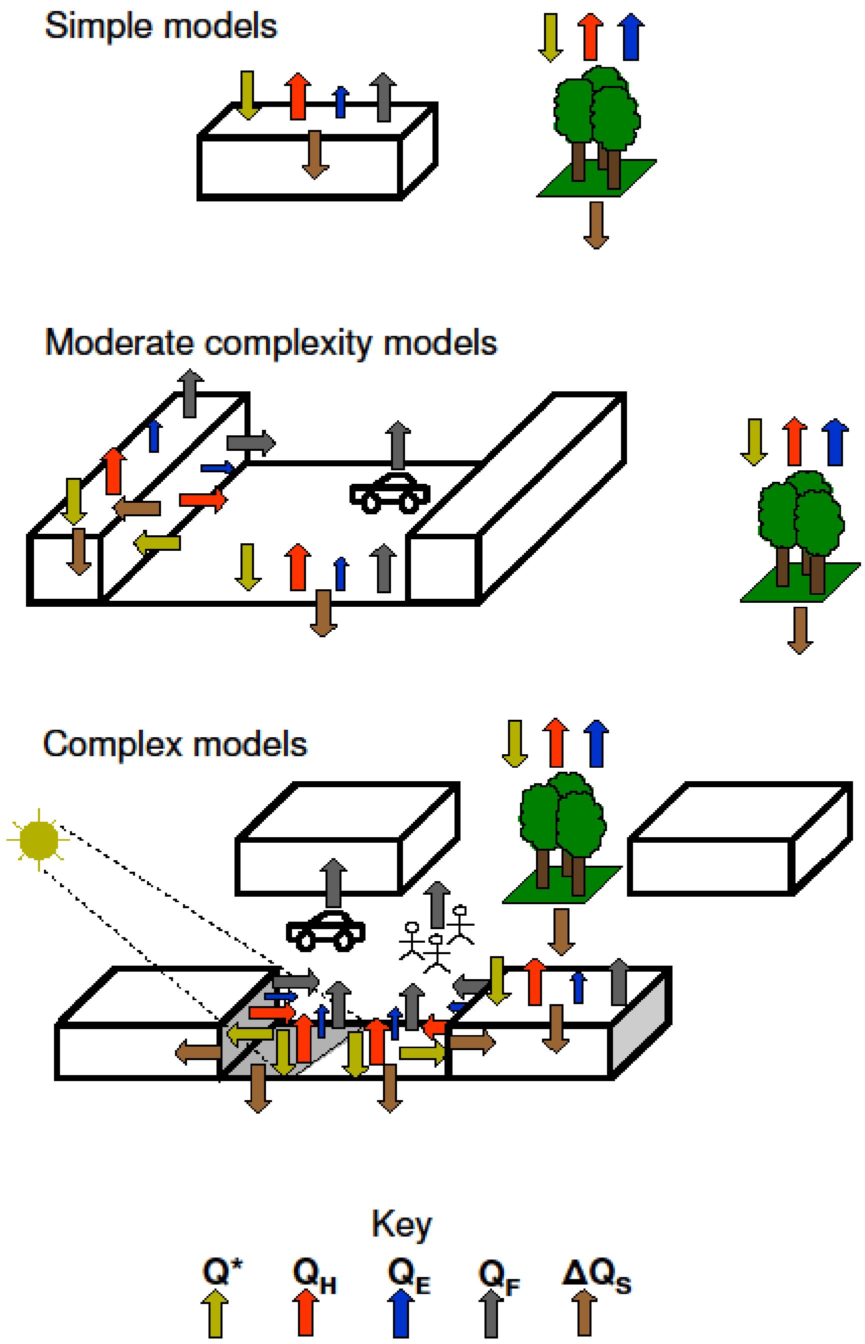

Figure 1: Conceptual figure of how surface energy balance exchanges are included in urban land surface models of different complexity. Note individual models have simple and complex features (Grimmond et al., 2011).

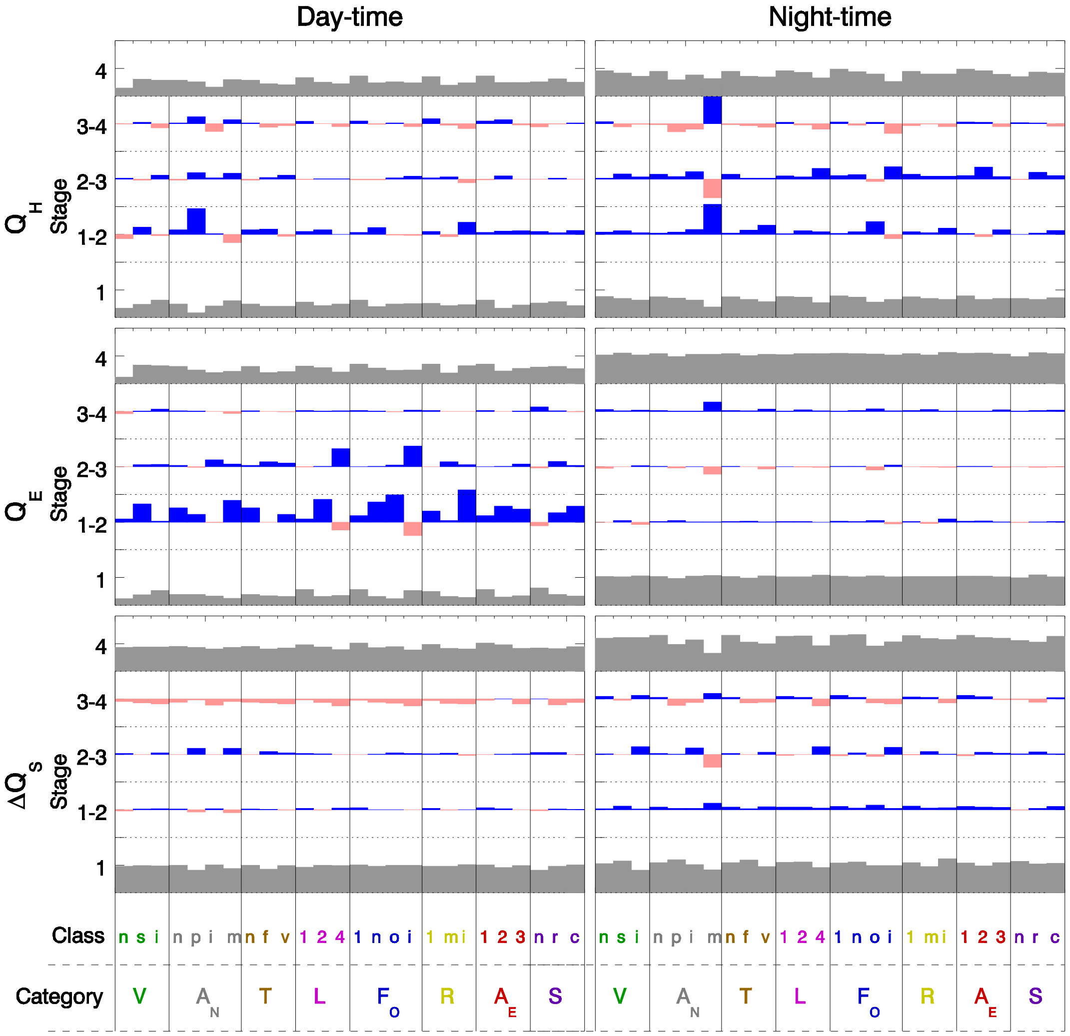

Figure 2: For each flux and physical category class (Table 4), the percentage of modelled data points within the specified observational errors (eqn. 1) for Stages 1 and 4 (grey) plus the change relative to the previous stage (eqn. 2; scaled between -100% and 100%, shown by the horizontal dotted lines). Blue shading indicates an improvement (> 0) and red degradation (< 0). Results are shown for day and night-time (with day defined as incoming solar radiation flux greater than 0 W m^{-2}). Codes definition for the physical categories and component classes (used in the x-axis) are given in Table 4.

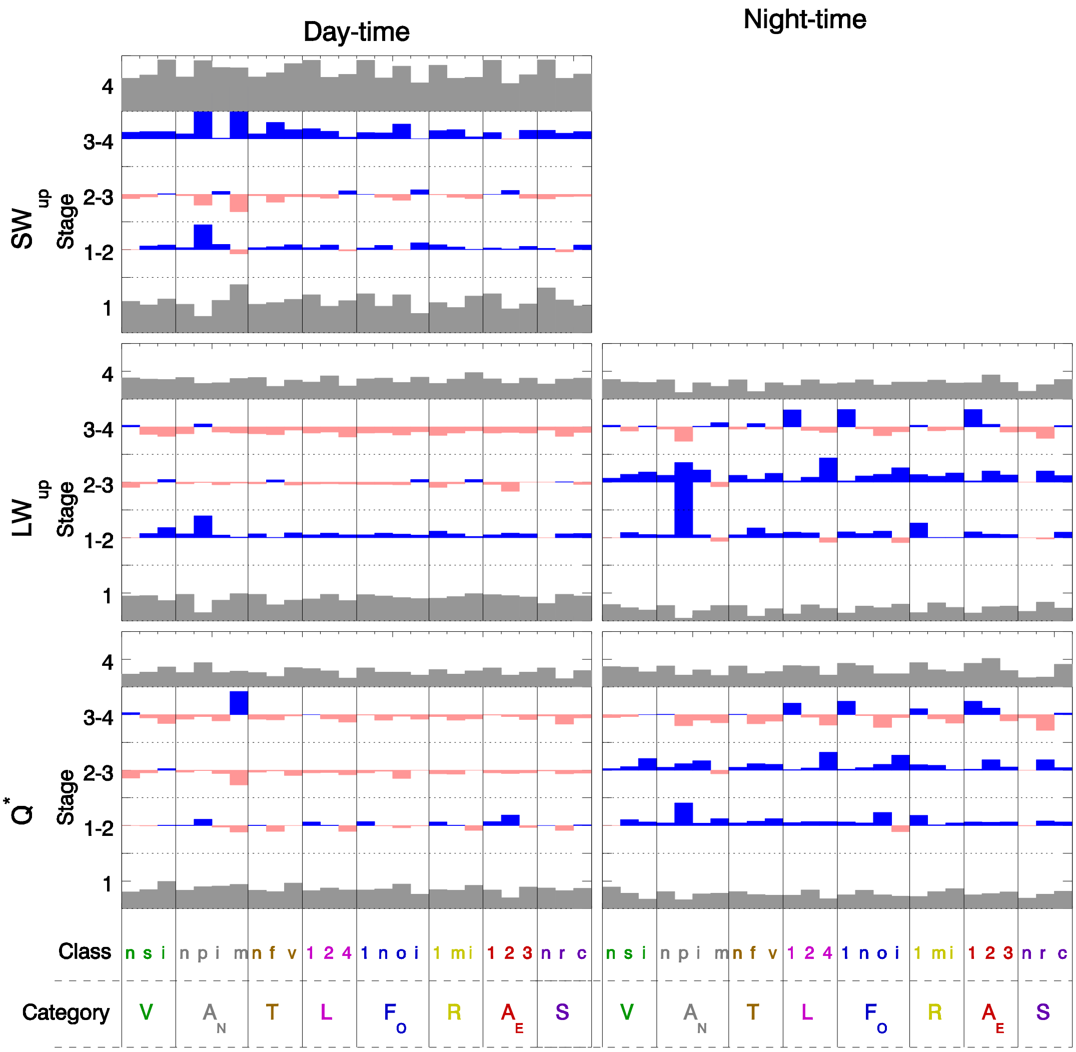

Figure 3: As for Fig. 2, but for the radiative fluxes.

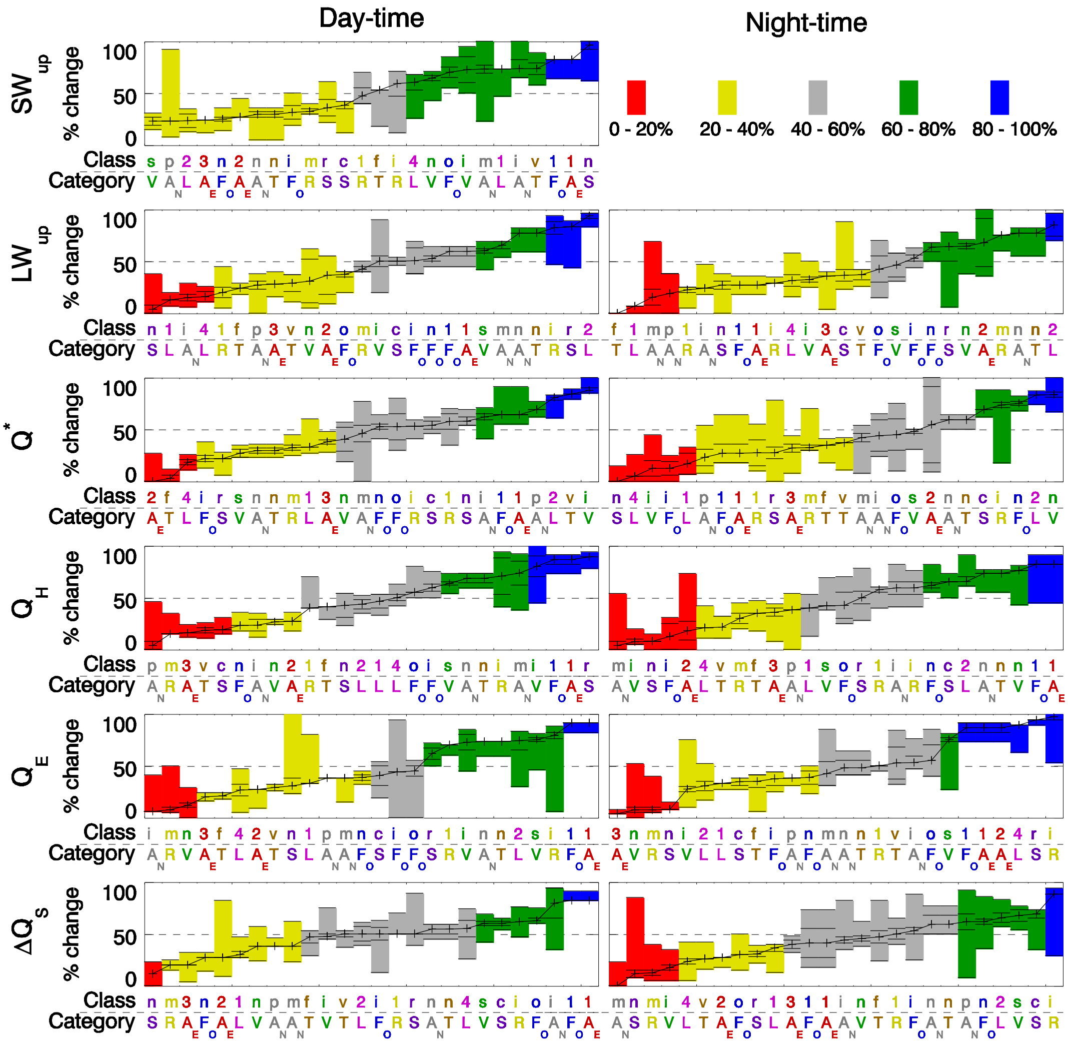

Figure 4: The subset of models within a class of a category improved compared to all models (P_{ca}, eqn. 4) ranked according to the median over the stages (for each flux, by time of day (as for Fig. 2)). Shading shows the range of results over the stages, with the individual results shown as horizontal lines within this. The colouring emphasises the values of the median over the stages, with 100% corresponding to all classes improved, 0% all classes degraded and 50% no change. Note X-axis code (Table 4) order changes between subplots because of ranking (Colour text is to aid differences to be noted).