Importance of initial state and atmospheric conditions for urban land surface models' performance

Article

Accepted Version

It is advisable to refer to the publisher's version if you intend to cite from the work. See Guidance on citing.

To link to this article DOI: http://dx.doi.org/10.1016/j.uclim.2013.10.006

Publisher: Elsevier

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

Importance of initial state and atmospheric conditions for urban land surface models' performance.

MJ Best [1,2], CSB Grimmond [2,3]

[1] Met Office, FitzRoy Road, Exeter, EX1 3PB, UK
[2] King's College London, Department of Geography, London, WC2R 2LS, UK
[3] Department of Meteorology, University of Reading, Earley Gate, PO Box 243, Reading, RG6 6BB, UK

Abstract

Urban land surface models (LSM) are commonly evaluated for short periods (a few weeks to months) because of limited observational data. This makes it difficult to distinguish the impact of initial conditions on model performance or to consider the response of a model to a range of possible atmospheric conditions. Drawing on results from the first urban LSM comparison, these two issues are considered. Assessment shows that the initial soil moisture has a substantial impact on the performance. Models initialised with soils that are too dry are not able to adjust their surface sensible and latent heat fluxes to realistic values until there is sufficient rainfall. Models initialised with too wet soils are not able to restrict their evaporation appropriately for periods in excess of a year. This has implications for short term evaluation studies and implies the need for soil moisture measurements to improve data assimilation and model initialisation. In contrast, initial conditions influencing the thermal storage have a much shorter adjustment timescale compared to soil moisture. Most models partition too much of the radiative energy at the surface into the sensible heat flux at the probable expense of the net storage heat flux.

Key Words: Energy balance, initial conditions, land surface models, soil moisture, urban, weather

1 Introduction

The resolution of weather forecasting models now enables land cover that was previously treated as sub-grid-scale heterogeneity to be explicitly resolved. This has resulted in a number of land surface schemes being developed to explicitly represent urban areas (e.g., Table 1). Combined with the increasing interest in the impact of climate change on the urban environment, this has led to the introduction of urban surfaces within the tiled land surface heterogeneity approach for climate models (e.g., Oleson et al. 2008a, McCarthy et al. 2010).

These urban land surface models (LSM) vary considerably in their complexity, from simple bulk representations, to models that divide a wall surface into several layers, each of which has its own surface energy balance (e.g., Table 1). To evaluate how these various urban LSM simulate the net all wave radiative flux and turbulent fluxes of sensible and latent heat, the models have been compared to data from observational campaigns (e.g., Ross and Oke, 1988, Grimmond and Oke, 1991, 2002, Masson et al., 2002, Best et al., 2006, Dupont and Mestayer, 2006, Hamdi and Schayes, 2007, Krayenhoff and Voogt, 2007, Oleson et al., 2008a, Kawai et al., 2009, Porson et al., 2009, Loridan et al., 2010, 2011, Järvi et al., 2011). Given the difficulties of taking observations within the urban environment, these campaigns have often been of relatively short duration of a few weeks to months (e.g., Grimmond et al., 1996, Oke et al., 1999, Voogt and Grimmond, 2000, Grimmond et al., 2004, Mestayer et al., 2005, Offerle et al., 2005, Newton et al., 2007). This has meant that it is not possible to determine the impact of the initial conditions, or the strategy chosen to spin-up the model from these initial conditions, on the performance of models.

Although the initial conditions can impact on a model integration, if the right spin-up strategy is adopted the influence of the initial conditions can be removed. To correctly spin-up the initial conditions, there needs to be a period of sufficient length that the memory of the initial conditions has been lost. For certain parts of the land system, such as deep soil moisture, the length of time required for spin-up can be many years, or even decades (e.g., Chen et al., 2007). For initial conditions to be correct, a dataset that spans multiple years prior to the period of interest are required. As this is usually unavailable, and would severely restrict possible study periods, spin-up strategies are often designed around repeating one given year of data a number of times. However, this strategy brings the model into equilibrium with that particular year rather than providing the correct initial conditions for the study period. As few urban observational datasets are multi-year, it is not possible to adopt this strategy for model spin-up, hence most urban studies have minimal spin-up and rely on their initial conditions being suitable. Thus the initial conditions can impact the evolution of the model, if incorrect values are used.

Many urban LSM were systematically assessed within the first urban LSM comparison (PILPS- urban) (Grimmond et al., 2010, 2011), using an observational period of 474 days. Initial results found a representation of vegetation is critical for accurate simulation of the surface turbulent fluxes of sensible and latent heat. This conclusion is supported when the LSM results are analysed at a higher temporal resolution which resolves the seasonal cycle (Best and Grimmond, 2013). However, it is well known that for vegetated surfaces, the partitioning between sensible and latent heat fluxes is largely determined by the availability of moisture in the soil and that changes to this soil moisture can occur on seasonal to annual timescales, especially for the moisture at deeper soil levels (e.g., Dirmeyer et al., 2006, Notaro et al., 2010). Therefore soil moisture initialisation can impact on the evolution of the sensible and latent heat fluxes for vegetation on these timescales. Hence it is reasonable to assume that soil moisture will impact the surface energy balance of an urban environment and that soil moisture profile initialisation could be important. Whilst urban models might have other water stores in their system (e.g., precipitation lying on a road or roof surface), the timescale associated with these stores does not compare in length to that of the deep soil moisture.

Urban environments typically have a higher thermal inertia compared to their surrounding rural areas, with net storage heat fluxes that are substantially larger than the comparative soil heat fluxes, causing a time delay in changes to the surface temperature of the urban facets (walls, roof, road) which will impact the energy balance fluxes (Grimmond and Oke, 1999). Any errors in the initial temperature profile within the

surface facets could persist, due to the large thermal inertia, impacting LSM performance. However the models use different representations of the surface, varying in complexity from a bulk representation of the urban area with a single surface temperature, to surface temperatures for a number of facets, to a number of surface temperatures for a single facet (such as several vertical levels for a wall). Thus it is difficult to identify a variable (e.g. surface temperature) that is observable at an appropriate scale that can be used to evaluate all of the models.

Assessing the impact and timescales that information is retained from initial conditions can be difficult, as often the data required to initialise can not be, or are not, observed. Unfortunately sensitivity studies can not distinguish between initialisation errors and model formulation errors. For instance, Best et al. (2006) demonstrated that the observed latent heat flux for an industrial site in Vancouver, Canada, could only be obtained if the model was initialised with unrealistically high values of soil moisture. Independent of horizontal advection possibly impacting these observation (Masson et al. 2002), it is possible to get realistic fluxes for the wrong reasons, such as unrealistic initial conditions.

So to distinguish between model formulation errors and initial condition, the simulation period needs to be sufficiently long so that the initial conditions are no longer retained while ensuring the seasonal variations in model performance are considered. In addition, observational datasets of short duration are unlikely to capture all of the synoptic weather conditions experienced at the site over each of the seasons, making it difficult to draw conclusions about the suitability of a model to represent an urban region in general. This suggests simulation for at least twelve months is required.

Whilst there are still few urban datasets of this length, the one used for the first international urban model comparison (Coutts et al., 2007a, b) is of sufficient length to allow for both the exploration of the impact of initial conditions on the performance of different urban models, and the impact of various atmospheric conditions on the performance of the urban models. Hence the objectives of this paper are two fold: first, to analyse urban LSM performance to assess the impact of user choices of initial conditions and the timescales these impacts persist, and second to assess how urban LSM performance changes with different weather conditions.

2 Methods

The urban model comparison was designed to provide a systematic evaluation of a number of urban LSMs. The overall goals of the comparison included the identification of: the dominant physical processes within the urban environment, the complexity required for an urban LSM to be suitable to represent the radiative, sensible and latent heat fluxes, and the subsequent parameter requirements for such a model. The model evaluations were performed for a single site using meteorological forcing data that was observed at the site provided to the modelling groups. The amount of information about the site (such as the vegetation fraction, the height to width ratio of the built environment, the buildings' thermal parameters, etc.) released to the modellers was provided in stages, so that analyses of the changes in results between stages could be used to address these goals (see details in Grimmond et al. 2011). To assess the impact of initial conditions "Phase 2/Stage 4" data, when all information about the observation site was provided, are analysed. Ideally the model performance should be its best at this stage, but Grimmond et al. (2011) found this is not the case for all models.

The forcing and evaluation data used were observed over a 16 month period (13 August 2003 to 28 November 2004) in Preston, a suburb in north Melbourne, Australia (Coutts et al., 2007a, b). The calculated land cover within the turbulent fluxes footprint was 44.5% buildings, 13% roads, 4.5% concrete, 15% grass, 22.5% other vegetation and 0.5% bare soil (Coutts et al. 2007a). The Oke (2004) Urban Climate Zone (UCZ) is 5 and the Loridan and Grimmond (2012a,b) Urban Zone for Energy exchange (UZE) is medium density (MD). The mean height to width ratio is 0.42 and the mean wall-to-plan ratio is 0.4.

The observations undertaken at 40 m, were analysed at 30 min intervals from samples at 1 Hz for the radiative fluxes and 10 Hz for the eddy covariance data. Further details are given in Coutts et al. (2007a, 2007b), Grimmond et al. (2011) and Best and Grimmond (2013). The anthropogenic heat flux was estimated following Sailor and Lu (2004) using inventory methods with the contributions from vehicles, building sector and human metabolism accounted for. The net storage heat flux, determined as the residual of the surface energy balance, accumulates the observational errors from the other terms in the energy balance.

Missing forcing data were gap filled to provide a continuous 30 min resolution data set. Evaluation is undertaken for only the time periods when all fluxes are available (8865 30-min periods), hence around 60% of the data are usable in this evaluation. To separate the impact of the initial conditions from the overall performance of the models, whilst also considering the seasonal cycle and enough data points per period, the analysis is performed for 60 day periods (Best and Grimmond, 2013). This yields seven 60 day periods, with the first and last being for the same time of the year. Prior to the first 60 day period there are an additional 53 days of forcing data. However, there are insufficient periods when all fluxes are available to allow meaningful statistics to be analysed from the model output. Thus, these 53 days are used as a spin-up period to try to remove the impact of the initial conditions and only the subsequent periods are analysed. No advice was given to the modelling groups to determine how they should set their initial conditions, or if an additional spin-up should be undertaken. These decisions were therefore made by the modelling groups themselves. This length of spin-up is likely to remove the initial conditions for moisture stores such as water on a road surface, but is not sufficient to spin-up the deeper soil moisture stores. Hence the results from the initial conditions impacting on model performance due to water availability are more likely to be connected to soil moisture and vegetation processes than the initial conditions for other water stores.

To assess the impact of the soil moisture initial conditions, the Bowen ratio is compared between the first and last of the 60 day periods. The Bowen ratio is explicitly linked to the soil moisture through the evapotranspiration from the vegetation fraction within the urban footprint. As some models were run by more than one modelling group (Table 1) with independent assumptions about the initial conditions, it is possible to separate the impact of the initial conditions from the evaluation of the model physics. The Bowen ratio, for the midday period (10:00 - 14:00 local standard time), is determined from the mean sensible and latent heat fluxes for the period. Similarly for the impact of other initial conditions, such as the temperatures of the various facets that make up the urban surface, the ratio of any two fluxes is determined from the

mean value of each flux over the same midday period.

To investigate the impact of atmospheric conditions, the observed and modelled net all-wave radiation (net shortwave plus net longwave), sensible and latent heat fluxes and net storage heat flux have been scaled by the incoming all-wave radiation flux so that a direct comparison can be made between these surface fluxes (Loridan and Grimmond 2012a,b). The median of these results, along with the standard error, has been calculated for a number of bins based upon the value of various atmospheric variables that are used to force the urban models. The size of the bins for each atmospheric variable have been chosen to allow the distribution of results across the range of the variable, whilst ensuring that there are enough data to create robust statistics for each bin. This results in a different number of bins between variable analysed ranging from eight for temperature to 16 for wind speed.

3 Results

3.1 Impact of model physics

A number of things influence the performance of each of the urban models; this includes: how the physics within the model are represented, the values of parameters used, the initial conditions of the prognostic variables in the model and any spin-up strategy that is used to reduce the influence of these initial conditions. Within the urban model comparison no spin-up protocol was specified. Therefore the initial conditions of each model could still influence model performance, especially as many urban model evaluation studies have only been able to use repeated used of the initial period rather than using a long period of preceding data due to the short duration of the observational datasets, alternatively spin-up is neglected. Although the choice of model parameters and the representation of the physical processes within the models can both influence performance they are treated here as "model physics" for the comparison to observations; and initial conditions are a second aspect.

The total observed precipitation during the simulation period was (all hours independent of availability of observed fluxes for evaluation) was 883 mm, which occurred during 1089 of the 30 min periods. These data were provided within the forcing data set. To assess the net variation in model performance the total cumulative evaporation (over all hours of the simulation) from each of the models are compared (Fig. 1a). It is evident that there is a large spread in the results with model 22 having a cumulative evaporative flux of 164 mm and model 44 of 821 mm (whole period). The evaporative flux for the restricted period when all the fluxes were observed was 214 mm (observed) whilst the two models were 52 and 358 mm respectively. These two models both have normalized Bowen ratios of $1.0 \leq \beta_N \leq 1.5$. Although the two models had similar evaporation rates for their first and last 60 day periods, their actual evaporation time series is quite different (Fig. 2). Model 22 underestimates the evaporation throughout the year, whereas model 44 has a much better fit to the observed latent heat flux, but in general overestimates during the winter period.

One model (50) has a net evaporation that is greater than the total precipitation over the period of the simulation (or integration—if not the whole period) and generally over estimates evaporation throughout the year. To sustain these large evaporation rates over an extended period of time would require irrigation. Unfortunately it is unknown which of the models simulated this. This data was not provided as part of the forcing data set and the area was under irrigation restrictions during the study period. So the net loss of water to the system simulated may be realistic over the period, and is unlike the results from any of the other models within the comparison (Fig. 1a). However it is likely that whilst the model physics has the largest impact on the differences in results, the initial conditions for model 50 are impacting its performance.

As each model does not change its representation of the physical processes during the simulation period, it is possible to identify the impact of the initial conditions on the performance for each of the models by comparing the behaviour of the model during the first and last 60 day period (i.e., the same period in consecutive years) relative to the observed behaviour at the same time. For example, an increase in the mean evaporative fraction (latent heat flux divided by the sum of the turbulent sensible and latent heat fluxes) in the observations over the same period of the year would suggest that there is more water available for evaporation in the second summer compared to the first. As the models are forced with the observed precipitation they should have the same trends as the observations despite any differences in their representation of the physical processes. So any disagreement between the model behaviour and observations is the impact of the initial conditions used for the simulation.

3.2 Initial Conditions: Bowen ratios
As the treatment of vegetation within the models plays a key role in their performance (Grimmond et al., 2010, 2011, Best and Grimmond, 2013), and evapotranspiration is influenced by soil moisture availability, the seasonal and annual evolution of soil moisture are important. Therefore one metric of the impact of the initial conditions is the change in the Bowen ratio between the start and end of the simulations.

The observations show that the seasonal cycle of the midday Bowen ratio (β) has an increase over the first 6 months of the summer period followed by a decrease in the winter time (Fig. 3a). During the summer months the larger values of net all wave radiation drive larger latent heat fluxes which reduce the soil moisture. This reduction in soil moisture is not sufficiently replenished by the precipitation that is reducing during this period, and hence evapotranspiration becomes limited by the availability of soil moisture. The net result is an increase in the average Bowen ratio (Fig. 3a). The observed Bowen ratio decreases after the larger autumnal rainfall in April and May. Overall, the first late spring (October/November) had a higher β than the following year which was wetter. The observed normalized midday Bowen ratio (first 60 day period normalised by the last 60 day period, β_N) was 1.24; i.e., β was larger (soil drier) at the start of the observation campaign than at the same time the following year (Fig. 1b).

The simulated Bowen ratio is influenced by a number of model processes which can include the soil water, the leaf area index of the vegetation, the depth of soil over which vegetation can extract water, the texture of the soils, etc. Thus initial soil moisture is not the only control on the relative performance of the various models but the comparison is informed by the assessment of the same period in two consecutive years for each model. The changes in the Bowen ratio between these two periods, removes differences in model physics within each model. Hence

comparing the normalised Bowen ratio of each model against that observed, allows us to assess how the change in moisture availability for the model differs to the observations.

Ideally the models would have the same proportional decrease in their Bowen ratios over the integration period, and hence the same normalised Bowen ratio as that observed. Six models do not simulate any latent heat flux (see Grimmond et al. 2010, 2011 for discussion) therefore they have an ill-defined β. Thus their β_N can not be compared. Of the remaining models, only 12 have a value of $1.0 \leq \beta_N \leq 1.5$ (i.e., close to the observed $\beta_N \pm 0.25$), whilst seven models have a β_N that is less than 1 (Fig. 1b). Eight models have $\beta_N > 1.5$, with three of these greater than 2.0 and a further two with $\beta_N > 2.5$. It is likely that this behaviour is caused by the initial conditions for soil moisture chosen by the user. The high β_N are probably related to low values of available water for evaporation at the start of the simulation restricting the latent heat flux and hence increasing the β, whereas the low values of β_N could be due to the initial available water being too high and hence not restricting the evaporation at all.

To demonstrate that the influence of initial conditions can be separated from the representation of model physics, the results for three models with similar cumulative evaporation, but different β_N behaviour (*viz*, models 41, 25 and 16) are shown in Figure 2. Although model 41 Bowen ratio ($1.0 \leq \beta_N \leq 1.5$) is similar between the first and last 60 day periods, the model clearly underestimates the evaporation during both summer periods (Fig. 2) suggesting the initial conditions are having little if any impact on the evolution of the model.

For model 25 ($\beta_N > 1.5$) the evaporation for the first summer period is underestimated whereas in the second summer period is in better agreement with the observations (Fig. 2). This suggests there is an available water limitation during the first summer but subsequently, after sufficient precipitation the model does not have this as an ongoing constraint so performs more realistically.

With model 16 ($\beta_N < 1.0$) the evaporation is more restricted during the second summer period compared than during the first, compared to the observations (Fig. 2). Given that the observed Bowen ratio decreases between the first and second summer periods (i.e., there is more water available for evaporation in the second summer that there is during the first), this suggests that whilst the physics of model 16 would generally underestimate the observed evaporation due to water availability. This indicates that the initial conditions have been set so that there is little limitation to the evaporation but through time the available water is depleted, despite receiving precipitation in the winter, so it becomes a constraint.

To consider the persistence and impact of the initial conditions three representative cases (Fig. 3a) are chosen: a) initial β more than double the final period value (Model 40); (b) starting β smaller than at the end (Model 50); and (c) similar β_N to the observations (Models 39, 46 with values of 1.22 and 1.23 respectively).

In the first case (40), the high β persists for the first 6 months when there is less seasonal rainfall, but in April/May when the rainfall increases again the β decreases to similar values to that observed (Fig. 3a). For the subsequent 60 day periods, the β are more similar to the observations, which suggests that model 40 was initialised with too dry soil moisture that restricted evapotranspiration unlike the observations. It is only once more rainfall occurs during the winter months, when the net radiation is limited, that the modelled soil moisture is able to recover to values which do not limit this evapotranspiration.

Although models 39 and 46 have β_N that are similar to the observations, the actual values of the β are not the same (Fig. 3a). Compared to the observations, model 39 has a similar shape to the seasonal cycle of β but there is an off-set throughout the simulation, with the model having higher values than the observations. The β values for model 46 compare well with the observations during the winter period, but have larger vales than observed in the autumn and both of the summer periods.

In the third case, model 50 initially does not follow the observed increase (December/January) but does start to increase in February/March (Fig. 3a). However, the β continues to increase in the autumn and winter months, before reducing in the spring. This may be caused by initial soil moisture that is too high. The initial reduction in soil moisture during the summer months is not sufficient to limit latent heat fluxes. However, the soil moisture continues to decrease during the autumn and winter which then restricts the evaporation and hence β increases, even as the rainfall start to increase. It is only once the accumulated rainfall reaches its larger values that the soil moisture is increased to a level that reverses this trend.

The differences in the evolution of the β between model 50 and others is not driven solely by the initial condition of soil moisture, but also by the amount of soil water that is removed via evapotranspiration. The average mean bias error (*MBE*) in the latent heat flux from model 50 is around double that from any of the other models with a positive bias (Table 2), and so it removes more soil moisture regardless of the initial conditions. Hence although the initial conditions of soil moisture are important, so is the relation between soil moisture and the limitation in evapotranspiration.

3.3 Initial Conditions: Model user assumptions

For two of the models in the comparison, results from identical configurations (including physical setup) were submitted by more than one group. This ensured that the only differences between the models were the initial conditions, ensuring a clearer analysis of the role of initial conditions on the evolution of the model fluxes, whilst removing differences due to the physical parameterisations. Fig. 3b shows the evolution of the β for these two models (now called "1", "2"). For both models, the simulations consist of results with an initial β substantially larger than the observed value ("dry") and a second set with β results in better agreement with the observations ("wet"). Although the timing of the reduction in the large β varies between the models, both simulations with initially large β have reduced values through the winter/spring

months, when the accumulated rainfall is higher. By the end of the simulation, all runs of the models have similar β which are close to the observed value. However, one of the model simulations (1_{wet}) has average midday β that is consistently lower than the observed value.

At the start 2_{wet} is closest to the observed β but there is substantial increase in the modelled β during February/March resulting in values that are close to the initialised dry run (2_{dry}) (Fig. 3b). This suggests that the soil moisture dries out to similar values despite the different initial conditions. However, this drying of soil moisture with subsequent increase in Bowen ratio is not seen in the observations. The initially wetter run (2_{wet}) stays drier during the period of larger accumulated rainfall compared to the initially drier model run (2_{dry}), suggesting that the representation of physics within the model is also sensitive to the evolution from its initial state and not just the initial conditions.

The impact of the initial conditions on 1_{dry} and 1_{wet} simulations can clearly be seen in Figure 3b. Whilst the two simulations show similar values of evaporation during the initial spin-up period (first 53 days), the evaporation from the 1_{dry} simulation is clearly restricted during the first summer period. This is caused by the onset of a restriction in water for evaporation that is due to the drier initial conditions set within this simulation. However, during the second summer period the evaporation rates are more similar between the two simulations, although 1_{wet} still has larger values than 1_{dry}. Hence although the impact of the initial conditions has been reduced by the second year, they have not be removed completely.

3.4 Initial Conditions: Sensible heat ratios

As building materials have high thermal inertia, the initial conditions for the temperature of the urban facets (walls, roof, road) could impact the simulations. In addition, although the soil thermal capacities are smaller than that for many building materials, the initial conditions for the soil temperatures in the models may also contribute. The net storage heat flux from the models does not distinguish between urban facets and the underlying soil, therefore these are considered together. If an impact from the initial surface and soil temperatures does exist, then we would expect a trend in the ratio of the net storage heat flux to net all wave radiation ($\Delta Q_S / Q^*$) between the start and end of the simulation period, and a reverse trend in turbulent sensible heat flux to net all wave ratio (Q_H / Q^*). Unlike β, there are no such trends in modelled ratios of the mean midday fluxes (e.g., Fig. 4), with the exception of Q_H / Q^* from model 50. There are some small trends in Q_H / Q^* in some cases (e.g., model 40, Fig. 4c) but not in $\Delta Q_S / Q^*$, even for model 50. Hence it is not obvious that the initial conditions chosen by the user for the soil temperatures or the surface temperature of each facet in their model have a significant impact on the simulation.

Although there are no obvious differences in the values of Q_H / Q^* and $\Delta Q_S / Q^*$ between the start and end periods, only five models (21, 22, 33, 44 and 49) have values of Q_H / Q^* which are consistently smaller than observed during the seasonal cycle (e.g. Fig. 4b). With the exception of model 22 which has a specified constant β (=1), these models all have an initial β that is lower than observed, and this persists throughout the seasonal cycle (not shown). The only other model with an initial β that is smaller than observed is model 50, which shows a trend in Q_H / Q^* over the year. Hence it is likely that these models were initialised with too high soil moisture. Also, apart from model 50, these models have values of Q_H / Q^* that are closer to the observed values by the end of the simulation than at the start. Hence this demonstrates that the initial conditions of soil moisture also impact on the sensible heat flux, through the surface energy balance, as might be expected. The implications of the initial conditions for the soil moisture are not so clear for the values of $\Delta Q_S / Q^*$ (e.g., Fig. 4b), although three of these models (21, 44 and 49) have values which are closer to those observed over the whole period than the other models.

There are a further four models (31, 32, 36 and 40) that have values of Q_H / Q^* which are lower than observed in the winter and spring period (e.g., Fig. 4c). Of these four models, three have an initial Bowen ratio that is substantially larger than observed, but are in much better agreement by the end of the simulation (not shown). However, these are not the only models that have an initial Bowen ratio that is substantially larger than that observed and hence the initial conditions on their own are not sufficient to explain the seasonal behaviour of Q_H / Q^* in these models. For the remaining majority of models the ratio of Q_H / Q^* is overestimated at all times of the year (e.g., Fig.4a), whilst the ratio of $\Delta Q_S / Q^*$ is underestimated, compared to the observations. This suggests that in general, the models put too much of the available net radiation at the surface into the sensible heat flux at the expense of the net storage heat flux at all times of the year. This result does not seem to depend upon the initial conditions, unless the soil moisture is set to values which are too high, resulting in a β that is lower than observed throughout the simulation.

This result can also be seen in the midday *MBE* for each of the fluxes (Table 2). For the majority of the models there is an overestimate in the sensible heat flux and an underestimate in both the latent heat flux and the net storage heat flux, including models 11 and 28 which have a large positive bias in the net radiation during the daytime. The average *MBE* for these fluxes are 50, -19 and -50 W m^{-2}, respectively; whilst the median model's *MBE* are 42, -11 and -50 W m^{-2}. Exceptions to this include some models that have a positive *MBE* in the latent heat flux, and consequently a negative *MBE* in the sensible heat flux (models 21, 33, 44 and 49). Model 50 has a positive bias in both the sensible and latent heat fluxes due to its excessive evaporation, whilst model 16 has a negative *MBE* for all fluxes, due to its large bias in the net radiation. Only models 22, 36 and 44 have positive *MBE* for the net storage heat flux in the afternoon, with all three models having a negative bias in the sensible heat flux.

3.5 Initial Conditions: Surface flux behaviour group by normalised Bowen ratio

To assess the change in performance of the models between the first and last 60 day periods, a Taylor plot (Taylor, 2001) is used for each flux (Fig. 5), in order to identify changes to three statistical measures within the same plot. Each statistic shown in the Taylor plot has been determined using the 30 min data for all models within each β_N category (section 3.1). The group of models with no latent heat flux are not impacted by the initial soil moisture conditions. This means that the change in performance between the first and last 60 day period for this group results from the differences in the climatological conditions and the model's response to these. As such, the results for the no evaporation

models can be used as a baseline to assess the impact of the initial conditions for the other groups of models. This is done by comparing the difference between the first and last 60 day periods for each group of models, to that from the group of models without evaporation. Note that models with no evaporation have no vegetation, but a model with no vegetation can still have evaporation if this process is included for other facets of the urban surface.

For net radiation, the best predicted flux, the models that have no latent heat flux have degradation in the root mean squared error (*RMSE*) and correlation at the end of the period compared to the start, although the normalised standard deviation remains about the same (Fig. 5a). This is in contrast to all of the other groups of models which show no real difference between the two periods in any of the statistics. This implies that the models generally show an improved performance in the net radiation at the end of the simulation compared to the start, due to the impact of initial conditions of soil moisture that offsets the degradation that results from the change in meteorological conditions.

For the turbulent sensible heat flux, the models that have no evaporation have a slight improvement in the *RMSE* and normalised standard deviation, but a slightly poorer correlation at the end of the simulation (Fig. 5b). The models with a $\beta_N < 1$ have degraded performance at the end of the period for all three of the statistical measures. However, these models have a substantial improvement in the latent heat flux for both the *RMSE* and correlation and an improvement to the normalised standard deviation resulting in good agreement with the observations at the end of the period (Fig. 5c). This is the only category of model that has a standard deviation in the latent heat flux for the first 60 day period which is larger than observed.

The models with $\beta_N > 1.5$ have a larger degradation in correlation between the last and first 60 day periods for the sensible heat flux compared to the models with no latent heat flux, but have improvements to the *RMSE* and particularly the normalised standard deviation that are larger than the models with no evaporation (Fig.5b). Those models in the $1.0 \leq \beta_N \leq 1.5$ class show an improvement in all of the statistics at the end of the period compared to the start. However, both β_N classes, plus the overall results for all models, have improved results in all of the statistical metrics for the latent heat flux at the end of the simulation compared to the start (Fig. 5c).

Notably the models with $\beta_N > 1.5$ have substantial improvements to the normalised standard deviation for both the sensible and latent heat flux (Figs. 5b, c). These models, initialised with a dry soil state, have a standard deviation that is too large for the sensible heat flux and too small for the latent heat flux during the first 60 day period. This is the expected behaviour if the vegetation was in an incorrectly water stressed state limiting the latent heat flux, resulting in too much energy partitioned into the sensible heat flux. The standard deviations for the last 60 day period do not agree with the observations, but are substantially closer for both fluxes.

All the model classes have an improved net storage heat flux at the end of the simulation compared to the start (Fig. 5d). There are some slight differences between the categories compared to the changes in the models with no evaporation, such as a larger improvement to the *RMSE* from the models with $\beta_N > 1.5$, but these differences are smaller than for the sensible and latent heat fluxes. Hence this shows that the initial conditions have more of an impact on the sensible and latent heat fluxes than on the net storage heat flux in general, due to the large influence that water availability has on the evaporative fraction.

For models that are influenced by soil moisture, there is a general improvement at the end of the simulation compared to the start for all fluxes. Whilst the precipitation over the integration period will influence these results, all models received the same precipitation input, so different changes in the errors for the various groups of models show that the initial conditions also contribute. The only exception to this is the sensible heat flux for the models that are initialised with too much soil moisture ($\beta_N < 1.0$, Fig. 5b). The largest improvements between the start and end of the simulation period are seen in the standard deviation of the latent heat flux for models initialised with either too much or too little soil moisture ($\beta_N < 1.0$ and $\beta_N > 1.0$ respectively). The models with too much initial soil moisture have more variability in the latent heat flux because it is not sufficiently constrained, whereas those with too little initial soil moisture have less variability due to an over constraint on the evaporation.

3.6 Meteorological forcing

As well as the initial conditions, it is important to know the effect of the meteorological forcing conditions on urban land surface model performance. To investigate this, the 30-min model results and observations are analysed across the range of each forcing variables (Fig. 6). The fluxes are scaled by the forcing incoming radiation ($Q\downarrow = K\downarrow + L\downarrow$, shortwave and longwave, respectively), with the mean and standard error shown for the bins from each meteorological forcing. These meteorological forcing (namely shortwave radiation, longwave radiation, rainfall, atmospheric temperature and specific humidity, wind speed and surface pressure) are the data that were given to each group in order to perform their integrations. As such, this data were consistent across all of the modelling groups. In addition, the models have also been analysed for the wind direction, which was not a meteorological forcing that was given to the modelling groups.

Whilst this analysis is not done according to particular weather types certain synoptic situations can be inferred, such as cyclonic conditions having lower shortwave radiation, higher longwave radiation, larger rainfall, higher wind speed and low pressures, or anticyclonic conditions having higher shortwave radiation, lower longwave radiation, less rainfall, higher temperature, lower wind speeds and high pressure.

The normalized Q_H and ΔQ_S observations have a similar magnitude response to atmospheric forcing variables, whereas Q_E is smaller. Figure 6 shows an example of this for a model with behaviour that is typical of the majority of models in the comparison. Q^* has the largest response to changes in variations of the forcing meteorological conditions. The almost linear trend between Q^* and the atmospheric temperature is probably linked directly to the radiative forcing of the surface and boundary layer (e.g., model 18, Fig. 6).

For both Q^* and Q_H the relation with the humidity forcing (Fig. 6) has two maxima and minima. Low values of specific humidity are most likely

related to hot, dry anticyclonic periods which have large insolation; whereas high specific humidity are linked to periods associated with precipitation events when the surface water is freely available, such as low pressure systems. Hence the two extremes of the humidity values lead to different behaviour in the net upward radiation and subsequently with sensible heat flux. For the wind speed, the observations have lower values of $Q*$ and Q_H and higher values of Q_E associated with high wind speeds and low pressure (Fig. 6), which is also consistent meteorological condition for low pressure systems. These synoptic conditions are likely associated with high humidity and rainfall.

In general the models capture the observed response to each of the atmospheric forcings (e.g., model 18, Fig. 6), including the non-linear behaviour with respect to the humidity and wind speed, and the double maximum and minimum for the specific humidity, which agrees well with the observations. However, the models do have a growing bias for increasing shortwave radiation forcing, with Q_H having an increasing positive bias and ΔQ_S with an increasing negative bias. The results for ΔQ_S show that the models have a flat response at larger values of the downward shortwave radiation forcing.

Whilst the majority of models reproduce the observed behaviour for each of the atmospheric forcing variables, the following have differing characteristics (Fig. 7). For Q_H, model 11 has higher values than observed for small and large downward shortwave and downward longwave radiation forcing (Fig. 7). The values are too high for small and especially large rainfall amounts, and there is a consistent positive bias to atmospheric temperature forcing. There are larger values than observed for lower atmospheric humidity, and an increasing positive bias for increasing wind speed. Finally, there are higher values than observed for low pressure. These results are all consistent with over estimating Q_H during low pressure synoptic storms.

For model 22, with a constant Bowen ratio specified, the results for Q_H do not capture the observed increase with larger downward shortwave radiation forcing (Fig. 7), whilst the trend for the downward longwave radiation is the reverse to that observed, apart from largest forcing values. There is a small response to atmospheric temperature, but the inverse to the observed trend, and a flat response to wind speed apart from a reduction at high wind speeds.

The Q_H for model 44 is underestimated compared to observations under large downward long wave radiative forcing (Fig. 7). During warmer temperatures the model results have the reverse trend to that observed, and decrease too rapidly with increasing wind speeds.

For model 50, the results for Q_H are too high compared to observations for downward shortwave radiation forcing (Fig. 7), whilst the modelled increasing trend in response to the downward long wave radiation between 300 and 400 W m^{-2} is not observed. The modelled values are too large for high rainfall amounts, whereas the increasing trend for wind speed between 5 and 15 m s^{-1} is not observed.

So the majority of the models can simulate the response of the observed surface fluxes to various atmospheric forcing conditions, but there is no obvious consistency between the minority that does not. These few models each have a different response which must be related to the details of their individual parameterisation schemes, rather than the representation of a generic physical process which they share with models in the majority that agree well with the observations.

4 Conclusions

For urban LSMs soil moisture initialisation can have a substantial impact on seasonal performance. If set too dry, urban LSMs are not able to adjust to realistic values until there is sufficient rainfall. Conversely, if initially too wet the model is not able to restrict the availability of water to vegetation. This may persist for a period that is longer than a year, but with smaller errors. This impact of the initial soil moisture state confirms the importance of vegetation and how this interacts with the available water store in the soil of urban areas.

The initialisation of soil moisture has implications for model evaluation studies. As many urban observational campaigns to date have been for short time periods (typically of the order of a few months), without well determined initial soil moisture conditions, evaluation of model performance may be hampered. Unfortunately, urban soil moisture data are still relatively rare. Thus, extreme care should be used to interpret model evaluation results if this is not explicitly addressed. For the modelling community to enhance their model evaluations and their model performance, future observational campaigns of fluxes in urban areas should be complemented with soil moisture observations.

The long term memory from soil moisture is more significant in the evolution of the simulations than for thermal storage, with the timescale for the adjustment of the initial conditions of surface temperatures being much shorter than for soil moisture and not apparent at the end of the simulation. However, the initial soil moisture can also impact on the sensible heat flux and the net storage heat flux throughout the simulation.

In general the models capture the observed trends in the surface fluxes for each of the atmospheric forcing variables, including the observed non-linear behaviour with respect to humidity and wind speed. However, the models are not able to capture the observed increasing trend in ΔQ_S with the larger downward shortwave radiation values. Overall, for increasing downward shortwave radiation, the models put too much energy into Q_H and too little into ΔQ_S. This is consistent with results from analysing the impact of initial conditions (section 3.3). Correcting such behaviour in the models requires future research, as it is not straight forward to increase ΔQ_S without having a similar impact on the turbulent sensible heat flux, due to their common interaction with the surface temperature which forms part of the thermal gradients that drive each of the fluxes.

Some models do not capture the observed relations with atmospheric forcing. However, no general conclusions can be drawn as these models do not share any common representation of their physical processes which implies that the results are not due to the representation itself. Rather it is probably the way in which these processes have been implemented within the parametrisations.

References

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

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

Best MJ, Pryor M, Clark DB, Rooney GG, Essery RLH, Ménard CB, Edwards JM, Hendry M.A, Porson A, Gedney N, Mercado LM, Sitch S, Blyth E, Boucher O, Cox PM, Grimmond CSB, Harding RJ (2011) The Joint UK Land Environment Simulator (JULES), Model description – Part 1: Energy and water fluxes. Geosci Model Dev 4: 677-699

Best MJ, Grimmond CSB (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

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

Chen F, Manning KW, Lemone MA, Trier SB, Alfieri JG, Roberts R, Tewari M, Niyogi D, Horst TW, Oncley SP, Basara JB, Blanken PD (2007) Description and evaluation of the characteristics of the NCAR high-resolution land data assimilation system. J. Appl. Meteorol. Climatol. 46: 694–713. doi: 10.1175/JAM2463.1

Coutts AM, Beringer J, Tapper NJ (2007a) Characteristics influencing the variability of urban CO2 fluxes in Melbourne, Australia. Atmos Environ 41: 51–62.

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

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

Dirmeyer PA, Gao X, Zhao M, Zhichong G, Oki T, Hanasaki N (2006) GSWP-2 multimodel analysis and implications for our perception of the land surface. Bull American Meteorol Soc 87: 1381-1397.

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

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

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

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

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

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

Grimmond CSB, Souch C, Hubble MD (1996) Influence of tree cover on summertime surface energy balance fluxes, San Gabriel Valley, Los Angeles. Clim Res 6: 45-57.

Grimmond CSB, Oke TR (2002) Turbulent heat fluxes in urban areas: observations and local-scale urban meteorological parameterization scheme (LUMPS). J Appl Meteorol 41: 792–810.

Grimmond CSB, Salmond JA, Oke TA, Offerle B, Lemonsu A (2004) Flux and turbulence measurements at a densely built-up site in Marseille: heat, mass (water and carbon dioxide), and momentum. J Geophys Res 109: D24101. doi:10.1029/2004JD004936.

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

Grimmond CSB, Blackett M, Best MJ, Baik J-J, Belcher SE, Beringer J, Bohnenstengel SI, Calmet I, Chen F, Coutts A, Dandou A, Fortuniak K, Gouvea ML, Hamdi R, Hendry M, Kanda M, Kawai T, Kawamoto Y, Kondo H, Krayenhoff ES, Lee S-H, Loridan T, Martilli A, Masson V, Miao S, Oleson K, Ooka R, Pigeon G, Porson A, Ryu Y-H, Salamanca F, Steeneveld GJ, Trombou M, Voogt JA, Young DT, Zhang N (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, Schayes G (2007) Validation of Martilli's urban boundary layer scheme with measurements from two mid-latitude European cities. Atmos Chem Phys 7: 4513–4526.

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

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

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

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

Järvi L, Grimmond CSB, Christen A (2011) The surface urban energy and water balance scheme (SUEWS): Evaluation in Los Angeles and Vancouver. J Hydrol 411: 219-237. doi:10.1016/j.jhydrol.2011.10.001

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

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

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

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

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

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

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

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

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

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

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

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

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

Loridan T Grimmond CSB 2012: Characterization of energy flux partitioning in urban environments: links with surface seasonal properties J. Appl. Meteorol. Climatol. 51,219-241 doi: 10.1175/JAMC-D-11-038.1

Loridan T, Grimmond CSB (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

Loridan T, Grimmond CSB, Grossman-Clarke S, Chen F, Tewari M, Manning K, Martilli A, Kusaka H, Best M (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, Grimmond CSB, Offerle BD, Young DT, Smith T, Järvi L, Lindberg F (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

Martilli A, Clappier A, Rotach MW (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, Grimmond CSB, Oke TR (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 MP, Best MJ, Betts RA (2010) Climate change in cities due to global warming and urban effects. Geophys Res Letters 37, L09705. doi:10.1029/2010GL042845

Mestayer PG, Durand P, Augustin P, Bastin S, Bonnefond J-M, Béenech B, Campistron B, Coppalle A, Delbarre H, Dousset B, Drobinski P, Druilhet A, Fréjafon E, Grimmond CSB, Groleau D, Irvine M, Kergomard C, Kermadi S, Lagouarde J-P, Lenonsu A., Lohou F, Long N, Masson V, Moppert C, Noilhan J, Offerle B, Oke TR, Pigeon G, Puygrenier V, Roberts S, Rosanti J-M, Saïd F, Salmond J, Talbaut M, Voogt J (2005) The urban boundary-layer field campaign in Marseille (UBL/CLU-ESCOMPTE) : Set-up and first results. Boundary-Layer Meteorol 114, 315-365.

Newton T, Oke TR, Grimmond CSB, Roth M (2007) The suburban energy balance in Miami, Florida. Geografiska Annaler Series A – Phys Geogr 89A: 331-347.

Notaro M, Liu Z, Gallimore RG, Williams JW, Gutzler DS, Collins S (2010) Complex seasonal cycle of ecohydrology in the Southwest United States. J Geophys Res 115: G04034. doi:10.1029/2010JG001382.

Offerle B, Grimmond CSB, Oke TR (2003) Parameterization of net allwave radiation for urban areas. J Appl Meteorol 42: 1157–1173.

Offerle B, Jonsson P, Eliasson I, Grimmond CSB (2005) Urban modification of the surface energy balance in the West African Sahel: Ouagadougou, Burkina Faso. J Clim 18: 3983-3995.

Oke TR (2004) Initial guidance to obtain representative meteorological observations at urban sites. Instruments and Observing Methods Report 81, WMO/TD 1250, 51 pp.

Oke TR, Spronken-Smith A, Jauregui E, Grimmond CSB (1999) The energy balance of central Mexico City during the dry season. Atmos Environ 33: 3919-3930.

Oleson KW, Bonan GB, Feddema J, Vertenstein M, Grimmond CSB (2008a) An urban parameterization for a global climate model: 1. Formulation and evaluation for two cities. J Appl Meteorol Climatol 47: 1038–1060.

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

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

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

Porson A, Clark PA, Harman IN, Best MJ, Belcher SE (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.

Ross SL, Oke TR (1988) Tests of three urban energy balance models. Boundary-Layer Meteorol 44: 73-96.

Ryu Y-H, Baik J-J, Lee S-H (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

Sailor DJ, Lu L (2004) A top-down methodology for developing diurnal and seasonal anthropogenic heating profiles for urban areas. Atmos Environ 38: 2737–2748.

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

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

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

Taylor KE (2001) Summarizing multiple aspects of model performance in a single diagram. J Geophys Res 106: 7183-7192.

Voogt JA, Grimmond CSB (2000) Modeling surface sensible heat flux using surface radiative temperatures in a simple urban area. J Appl Meteorol. 39: 1679–1699.

Table 1: Models (name, versions (V) and user groups (G)) in Phase 2/Stage 4 of the first international urban model comparison. Model results are presented with anonymous numerical code.

Model name	V/G	References
Building effect parameterization (BEP)	1/1	Martilli et al. (2002)
BEP coupled with building energy model	1/1	Martilli et al. (2002); Salamanca et al. (2009, 2010) ; Salamanca and Martilli (2010)
Community Land Model – urban (CLM-urban)	1/1	Oleson et al. (2008a, 2008b)
Institute of Industrial Science urban canopy model	1/1	Kawamoto and Ooka (2006, 2009a, 2009b)
Joint UK land environment simulator (JULES)	4/2	Essery et al. (2003); Best (2005); Best et al. (2006); Best et al. (2011)
Local-scale urban meteorological parameterization scheme (LUMPS)	2/1	Grimmond and Oke (2002); Offerle et al. (2003); Loridan et al. (2011)
Met Office Reading urban surface exchange scheme (MORUSES)	3/1	Harman et al. (2004a, 2004b); Porson et al. (2010)
Multi-layer urban canopy model	1/1	Kondo and Liu (1998); Kondo et al. (2005)
Nanjing University urban canopy model-single layer	1/1	Masson (2000); Kusaka et al. (2001)
National and Kapodistrian University of Athens model	1/1	Dandou et al. (2005)
Noah land surface model/single-layer urban canopy model	3/3	Kusaka et al. (2001); Chen et al. (2004); Loridan et al. (2010)
Seoul National University urban canopy model	1/1	Ryu et al. (2011)
Simple urban energy balance model for mesoscale simulation	1/1	Kanda et al. (2005a, 2005b); Kawai et al. (2007, 2009)
Single column Reading urban model tile version	4/1	Harman and Belcher (2006)
Slab urban energy balance model	1/1	Fortuniak (2003); Fortuniak et al. (2004, 2005)
Soil model for submesoscales (urbanized)	1/1	Duport and Mestayer (2006); Dupont et al. (2006)
Temperatures of urban facets (TUF) 2D	1/1	Krayenhoff and Voogt (2007)
Temperatures of urban facets (TUF) 3D	1/1	Krayenhoff and Voogt (2007)
Town energy balance (TEB)	1/1	Masson (2000); Masson et al. (2002); Lemonsu et al. (2004); Pigeon et al. (2008)
Town energy balance (TEB) with multi-layer option	1/1	Hamdi and Masson (2008)
Vegetated urban canopy model	1/1	Lee and Park (2008)

Table 2: Mean bias error (W m^{-2}) for each flux over the day (24 h) and for midday period (10:00 – 14:00), for each model and for the average and median of all of the models, over the seven 60 day periods. (day 54 to day 473)

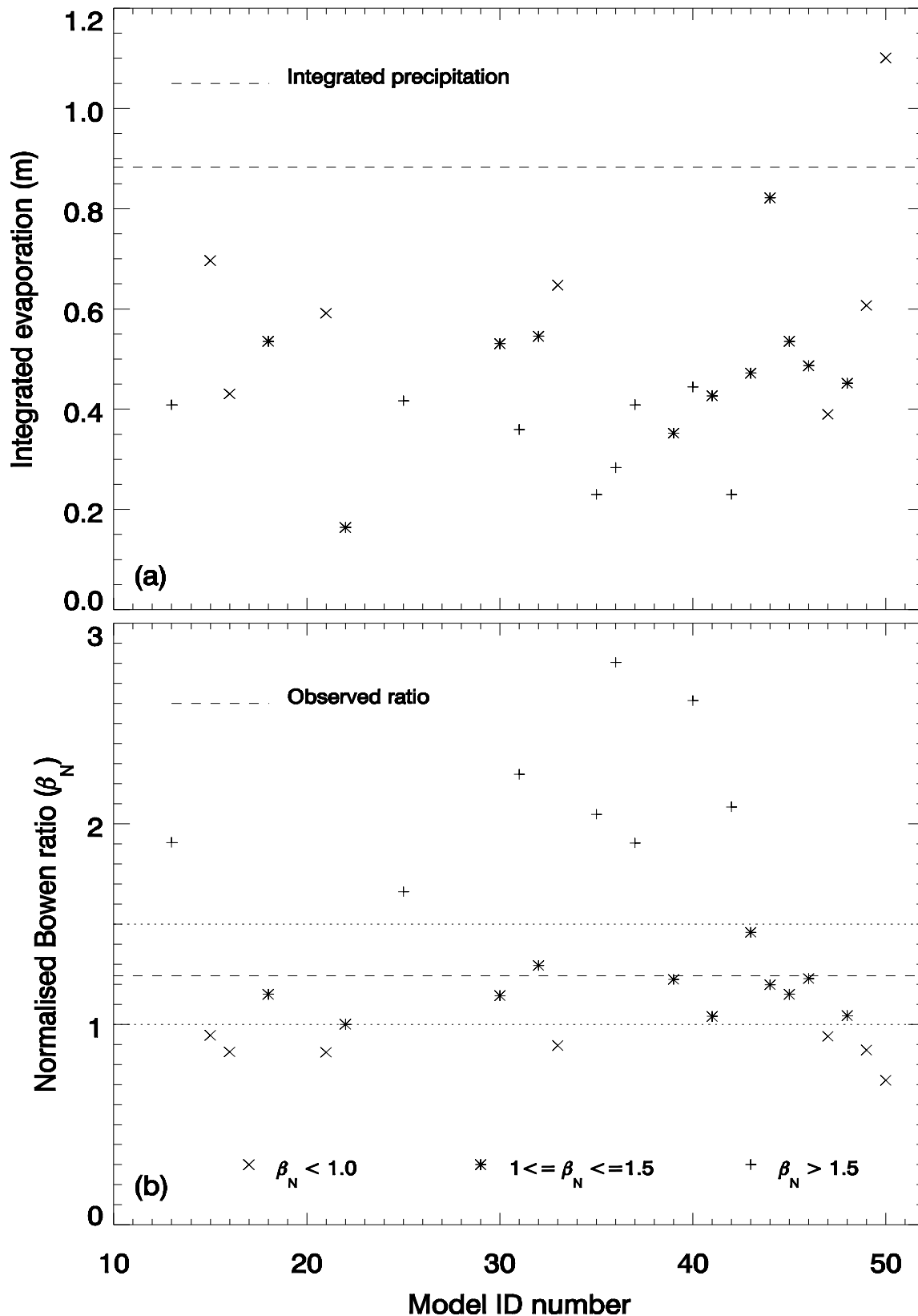| Model ID	Q*		Q_H		Q_E		ΔQ_S	
	24 h	Midday	24 h	Midday	24 h	Midday	24 h	Midday
11	10.5	110.7	70.2	213.2	-34.5	-80.7	-35.8	-33.6
12	-9.6	-37.5	34.9	119.5	-34.5	-80.7	-20.5	-88.1
13	-13.7	-36.6	0.6	20.2	-7.1	-18.8	-17.7	-49.7
14	-9.0	-34.7	35.5	118.6	-34.5	-80.7	-20.5	-84.3
15	-7.7	-9.1	0.5	88.2	16.2	34.5	-13.0	-96.1
16	-23.0	-92.9	-10.5	-0.4	-3.2	-10.8	-19.9	-93.5
18	2.3	-8.3	10.0	60.5	0.5	0.5	-18.8	-81.0
20	-4.7	13.6	55.5	131.2	-34.5	-80.7	-36.2	-48.7
21	-7.0	-18.2	-5.4	-30.1	4.7	12.8	-16.7	-12.7
22	-4.6	2.8	-0.7	-41.6	-26.2	-59.6	11.8	92.3
25	0.3	-13.1	25.8	72.4	-8.8	-10.0	-17.1	-73.5
28	18.7	113.4	59.3	213.8	-34.5	-80.7	-16.6	-31.4
30	3.2	-14.1	12.2	79.6	0.4	0.6	-19.8	-106.0
31	-2.6	-17.5	3.9	39.1	-11.7	-19.9	-17.2	-80.3
32	-0.8	-20.6	11.2	19.5	-2.2	-0.3	-20.3	-51.5
33	-6.2	-21.7	-12.4	-25.7	10.0	35.9	-14.4	-45.7
35	-27.8	-47.9	13.6	1.5	-19.2	-52.4	-32.6	-8.8
36	-8.6	-5.1	19.2	-1.1	-16.5	-33.2	-14.7	27.2
37	-13.8	-35.6	0.4	18.5	-7.1	-18.6	-17.6	-47.2
38	3.1	16.2	44.7	127.4	-34.5	-80.7	-17.6	-42.3
39	-14.7	-45.5	1.8	42.3	-9.2	-20.4	-17.7	-79.2
40	-3.4	-27.0	13.2	53.2	-8.1	-9.3	-18.8	-82.1
41	-10.9	-13.0	8.4	35.6	-9.8	-15.8	-20.0	-44.4
42	-29.8	-48.3	20.7	5.8	-19.2	-52.5	-41.7	-13.3
43	-10.1	-31.2	1.7	21.9	-3.1	7.6	-19.1	-72.5
44	3.7	11.6	-24.4	-44.4	23.4	36.3	-5.8	8.0
45	2.3	-8.3	10.0	60.5	0.5	0.5	-18.8	-81.0
46	-4.4	-20.7	9.0	53.8	-2.9	-0.5	-21.0	-85.9
47	-8.6	-18.9	11.1	41.8	-8.8	-9.6	-21.4	-62.9
48	5.4	-7.6	16.7	38.9	-7.6	-15.1	-14.1	-43.1
49	-6.9	-16.9	-6.6	-32.5	5.6	13.4	-16.4	-9.5
50	-6.6	-8.2	10.6	88.9	38.4	83.9	-38.5	-87.7
Average	-5.8	-12.2	13.8	49.7	-8.7	-18.9	-19.6	-50.3
Median	-6.2	-16.9	10.6	41.8	-7.6	-10.8	-18.8	-49.7

Fig. 1: Individual model values for (a) total evaporation (m) simulated for the period and total cumulative precipiptation observed (dashed line), and (b) normalised Bowen ratio (first 60 day period normalised by the last 60 day period) relative to the observations (dashed line). To preserve anonymity each model is identified by a random number (between 10 and 50). Symbols in both plots are based on the normalized Bowen ratio (see key) and dotted lines in (b).
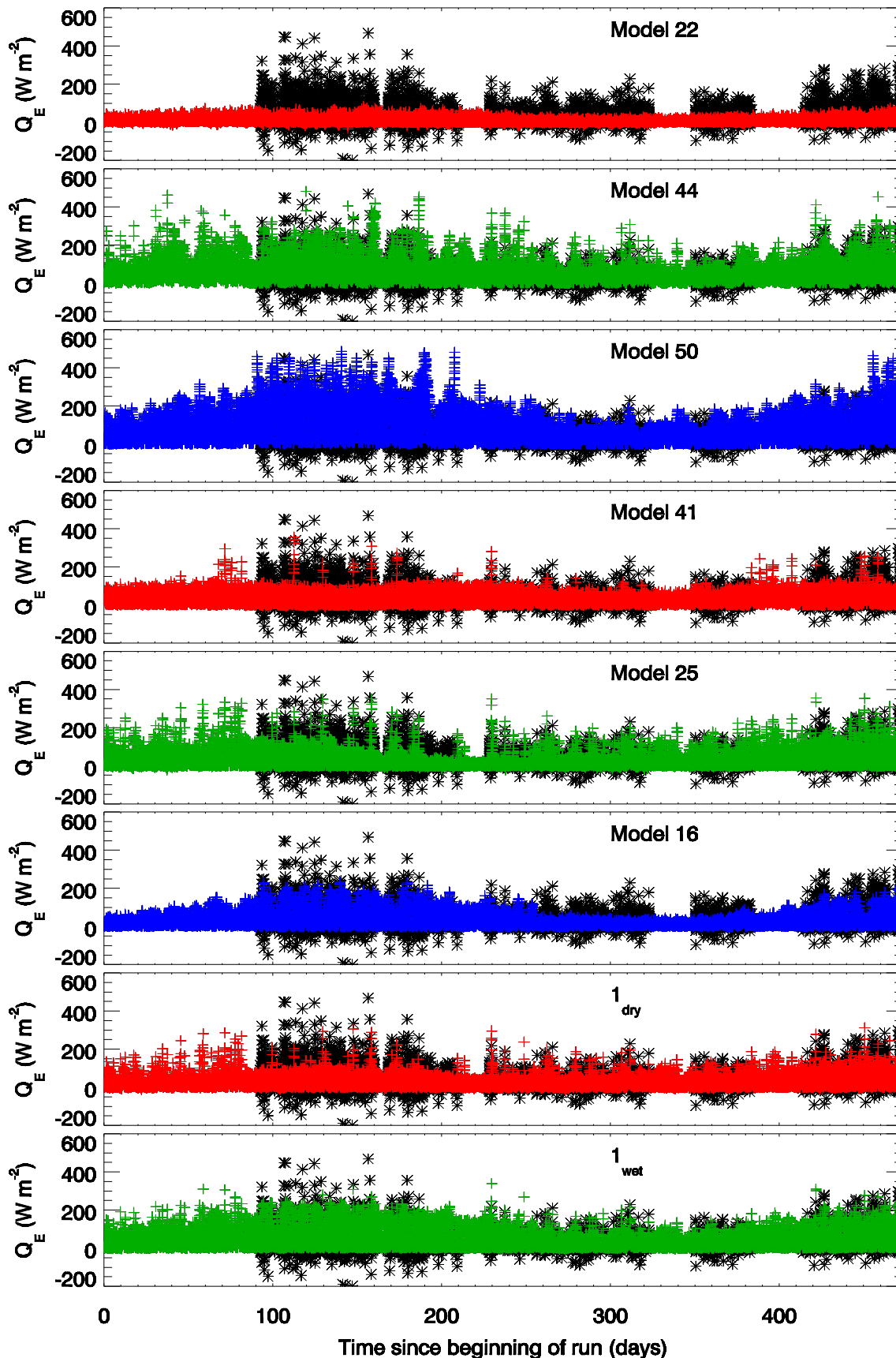
Fig 2: Time series of observed and modelled latent heat flux for a selection of (Black asterisk – observations, coloured pluses – model)

Fig. 3: Evolution over the seasonal cycle of the Bowen ratios calculated from the mean midday (10:00 – 14:00) fluxes for (a) a selection of models (see text) (b) two urban LSM ("1", "2") each run by two separate groups.
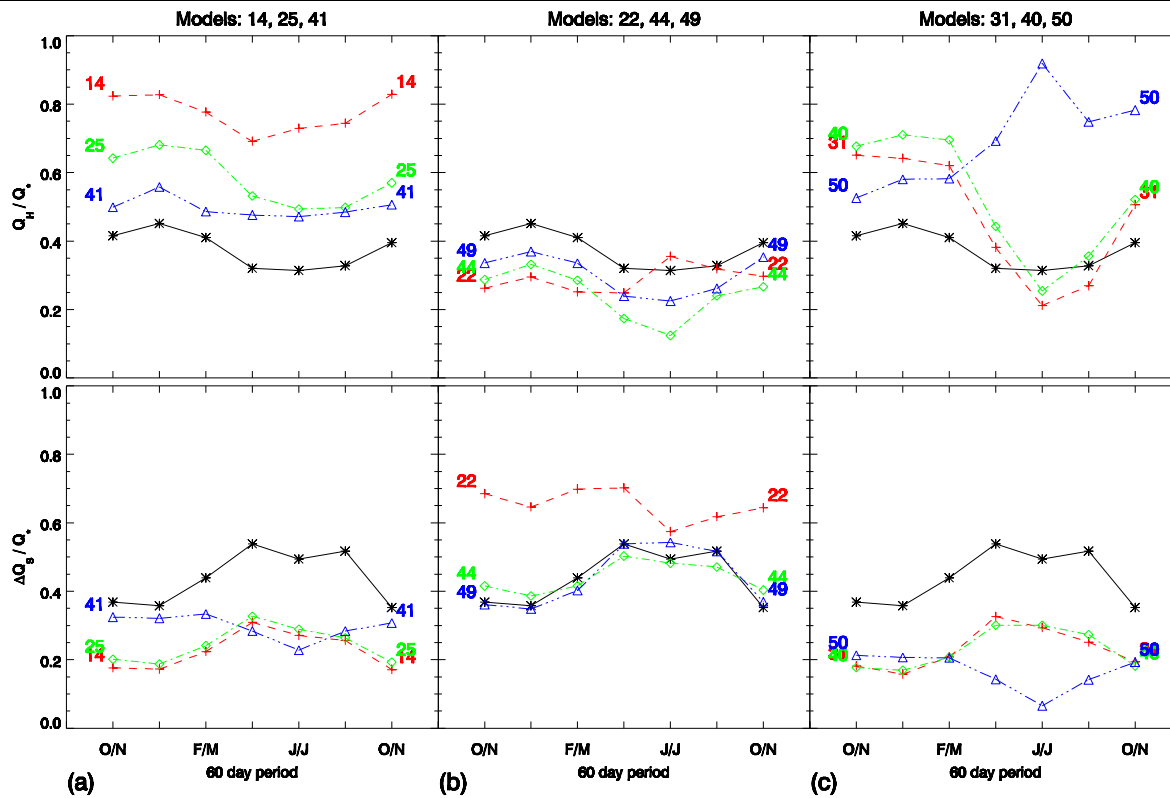
Fig. 4. Observed (black line *) and modelled ratio of the average midday (10:00 – 14:00) (top) turbulent sensible heat flux to the net all wave radiation flux (Q_H/Q^*) and (bottom) net storage heat flux to the net radiation flux ($\Delta Q_S/Q^*$) over the seasonal cycle for models with: (a) typical performance, (b) Q_H/Q^* smaller than observed, and (c) Q_H/Q^* smaller than observed in winter and spring along with model 50.
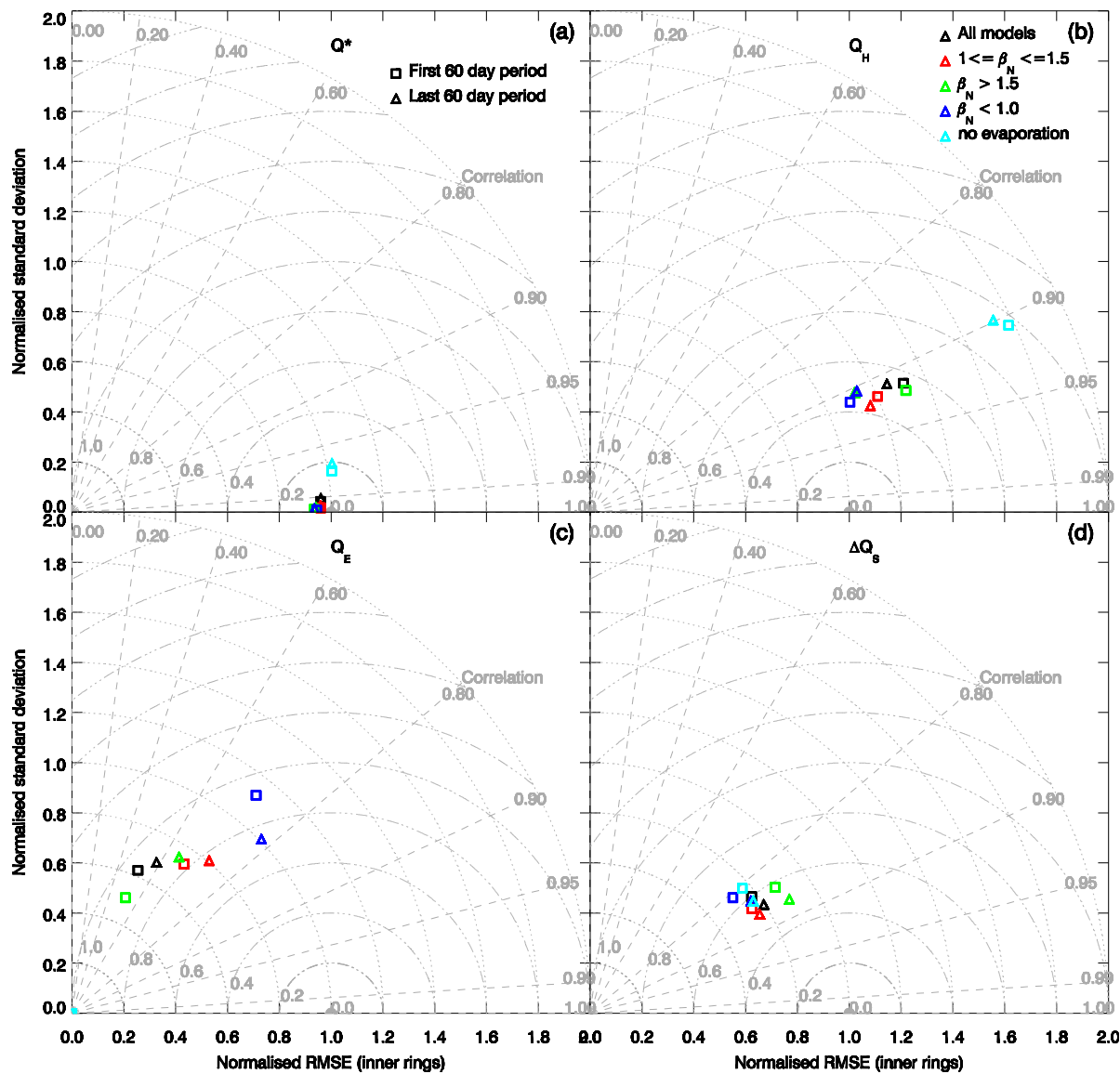
Fig. 5: Normalised Taylor plot for first (square) and last (triangle) 60 day periods for (a) net all wave radiation, (b) turbulent sensible and (c) latent heat and (d) net storage heat fluxes. Taylor plots have the correlation coefficient on the polar axis, the normalised standard deviation on the radial y-axis and the normalised *RMSE* (x-axis) on the internal circular axes (Taylor 2001). Performance for the average of all the models, plus the average of the models grouped by their normalised Bowen ratio (β_N), as identified in Fig. 2 (see key). For (c) models with no evaporation the points plot at the origin.
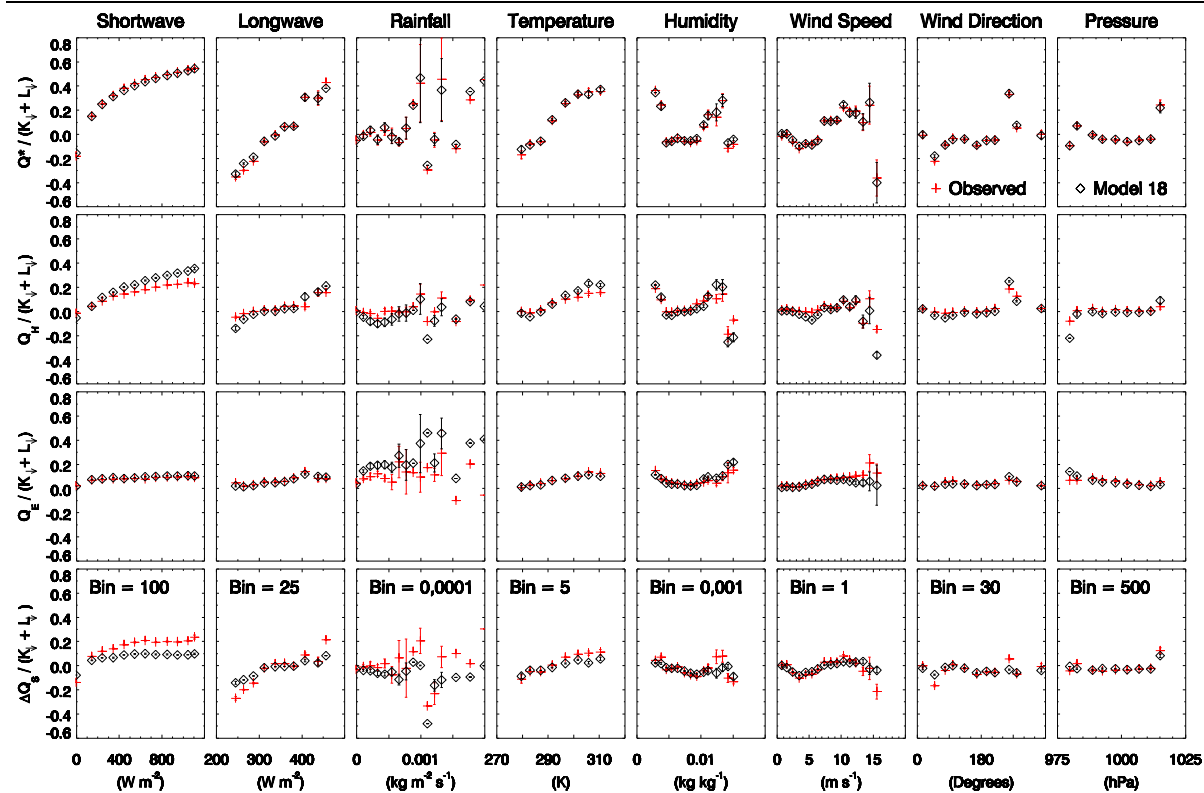
Fig. 6: Observed and a typical model (model 18) fluxes normalized by observed incoming all wave radiation ($Q\downarrow = K\downarrow + L\downarrow$) binned (bin sizes as indicated on each figure, units as per figure) by atmospheric forcing type: median (symbol) and standard error (vertical bar)
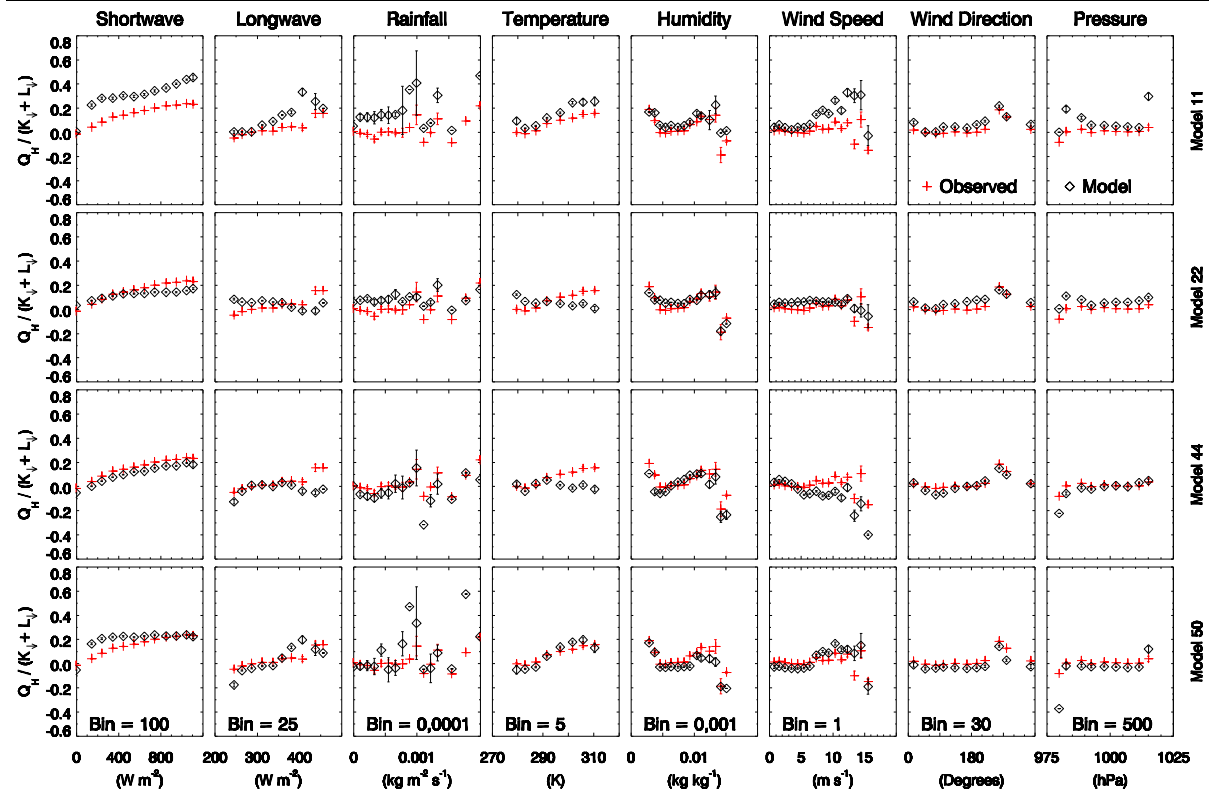
Fig. 7: Sensible heat flux normalized by incoming all wave radiation ($Q\downarrow = K\downarrow + L\downarrow$) binned (bin sizes as indicated on each figure, units as per figure) by atmospheric forcing type (columns) for four models (11, 22, 44, 50 rows) with characteristics that differ from the observations: median (symbol) and standard error (vertical bar)