# The water balance representation in Urban-PLUMBER land surface models

**H.J. Jongen[1,2], M. Lipson[3], A.J. Teuling[1], C.S.B Grimmond[4], J.-J. Baik[5], M. Best[6], M. Demuzere[7,8], K. Fortuniak[9], Y. Huang[10], M.G. De Kauwe[11], R. Li[12,13], J. McNorton[14], N. Meili[15,16], K. Oleson[17], S.-B. Park[18], T. Sun[12], A. Tsiringakis[2,19], M. Varentsov[20], C. Wang[10,21], Z.-H. Wang[22] and G.J. Steeneveld[2]**

[1]Hydrology and Environmental Hydraulics, Wageningen University, Wageningen, The Netherlands.
[2]Meteorology and Air Quality, Wageningen University, Wageningen, The Netherlands.
[3]Bureau of Meteorology, Canberra, Australia.
[4]Department of Meteorology, University of Reading, Reading, United Kingdom.
[5]School of Earth and Environmental Sciences, Seoul National University, Seoul, South Korea.
[6]Met Office, Exeter, United Kingdom.
[7]Urban Climatology Group, Department of Geography, Ruhr-University Bochum, Bochum, Germany.
[8]B-Kode, Ghent, Belgium.
[9]Department of Meteorology and Climatology, Faculty of Geographical Sciences, University of Łódź, Łódź, Poland.
[10]School of Meteorology, University of Oklahoma, Norman, Oklahoma, United States of America.
[11]School of Biological Sciences, University of Bristol, Bristol, United Kingdom.
[12]Institute for Risk and Disaster Reduction, University College London, London, United Kingdom.
[13]Department of Hydraulic Engineering, Tsinghua University, Beijing, China.
[14]European Centre for Medium-Range Weather Forecasts (ECMWF), Reading, United Kingdom.
[15]Department of Civil and Environmental Engineering, National University of Singapore, Singapore, Singapore.
[16]Future Cities Laboratory Global, Singapore-ETH Centre, Singapore, Singapore.
[17]U.S. National Science Foundation National Center for Atmospheric Research (NSF NCAR), Boulder, Colorado, United States of America.
[18]School of Environmental Engineering, University of Seoul, Seoul, South Korea.
[19]European Centre for Medium-Range Weather Forecasts (ECMWF), Bonn, Germany.
[20]Faculty of Geography/Research Computing Center, Lomonosov Moscow State University, Moscow, Russia.
[21]Department of Geography and Environmental Sustainability, University of Oklahoma, Norman, Oklahoma, United States of America.
[22]School of Sustainable Engineering and the Built Environment, Arizona State University, Tempe, Arizona, United States of America

## Key Points:

- We evaluate the water balance in 19 urban land surface models (ULSM) from the Urban-PLUMBER project.
- ULSMs capture the timing of water fluxes more accurately than their magnitude.
- The water balance appears unclosed in 43% of the model runs (19 models at 20 sites).

Corresponding author: Harro Jongen, harro.jongen@wur.nl

**Abstract**

Urban Land Surface Models (ULSMs) simulate energy and water exchanges between the urban surface and atmosphere. When part of numerical weather prediction, ULSMs provide a lower boundary for the atmosphere and improve the applicability of model results in the urban environment compared with non-urban land surface models. However, earlier systematic ULSM comparison projects assessed the energy balance but ignored the water balance which is coupled to the energy balance. Here, we analyze the water balance representation in 19 ULSMs participating in the Urban-PLUMBER project using results for 20 sites spread across a range of climates and urban form characteristics. As observations for most water fluxes are unavailable, we examine the water balance closure, flux timing, and magnitude with a score derived from seven indicators. We find that the water budget is only closed in 57% of the model-site combinations assuming closure when annual total incoming fluxes (precipitation and irrigation) fluxes are within 3% of the outgoing (all other) fluxes. Results show the timing is better captured than magnitude. No ULSM has passed all good water balance indicators for any site. Our results indicate models could be improved by explicitly verifying water balance closure and revising runoff parameterizations. By expanding ULSM evaluation to the water balance and related to latent heat flux performance, we demonstrate the benefits of evaluating processes with direct feedback mechanisms to the processes of interest.

## Plain Language Summary

Urban environments have their own local climates including typically higher nocturnal temperatures compared with rural areas. Ideally, modeling cities should capture their influences on the atmosphere above them. As the energy and water balances are linked by evaporation, a good water balance representation will support a good energy balance simulation. Focusing on the water balance, we find the water balance in models could be improved by paying attention to closure and runoff.

## 1 Introduction

The impact of urbanization on the local climate and hydrology has sparked scientists' interest and inspired research for centuries (e.g. Howard, 1833; Oke, 1982; Fletcher et al., 2013; Hamdi et al., 2020). With the increasing population in cities (United Nations, 2018) more people are impacted by increased heat stress and flooding (Heaviside et al., 2016; Gasparrini et al., 2017; Zhou et al., 2019; Botzen et al., 2020). Spatial morphological heterogeneity and human interactions make understanding the urban climate challenging (Kotthaus & Grimmond, 2014a; Sun et al., 2018; Koopmans et al., 2020; Demuzere et al., 2022), but weather and climate models need to include the effects of urban areas, as they locally exacerbate extreme events (Oleson et al., 2008; Ronda et al., 2017; Hertwig et al., 2020). Examples are increased flooding due to high impervious fractions (Zhou et al., 2019) and increased heat stress during heat waves resulting from the high heat storage capacity (Lemonsu et al., 2015). Therefore, models need to capture the impact of urban areas on their climate.

Researchers have developed, evaluated, and improved Urban Land Surface Models (ULSMs) simulating the interaction of the urban surface with the atmosphere. Coupled with a numerical weather prediction or climate model, ULSMs serve as a lower boundary condition and improve the model performance for urban environments (Tewari et al., 2007). ULSMs make different simplifying assumptions regarding urban geometry: a single homogeneous, impervious slab; multiple, individually homogeneous slabs; two-dimensional canyons; or 3D streets with individual buildings (Grimmond et al., 2009). These models also differ in whether and how they include physical processes like anthropogenic heat, irrigation, and snow processes (Lipson et al., 2023a). To evaluate their performance, these models are compared with observations (e.g. Ross & Oke, 1988; Grimmond & Oke, 2002; Hamdi & Schayes, 2007; Krayenhoff & Voogt, 2007; Porson et al., 2010). Although these individual evaluations were sometimes based on the same observations (Grimmond et al., 2009), the lack of a systematic approach prevented consistent comparison of the schemes. To compare the wide variety of models, two successive comparison projects applied a systematic approach. The first systematic comparison of ULSMs generally followed the PILPS protocol (project for intercomparison of land surface parameterization schemes, Henderson-Sellers et al. (1996)), hence PILPS-Urban (Grimmond et al., 2010, 2011). Individual modelers received meteorological input and surface characteristics to enable them to run their models. In total, 32 models completed simulations for a site in Vancouver and one in Melbourne. Grimmond et al. (2011) concluded that increased model complexity did not necessarily benefit model performance.

The second intercomparison, Urban-PLUMBER (Lipson et al., 2023a), assesses 30 models initially at the PILPS-Urban Melbourne site and adopts benchmarks following the PLUMBER project (Best et al., 2015). Benchmarks serve as a relative reference, to which models are compared to assess whether a cohort performs better (or not) than the benchmark and if input information is utilized effectively. Urban-PLUMBER is extended to the 20 sites presented by Lipson et al. (2022a) in the second phase (Lipson et al., 2023b). The Urban-PLUMBER models outperform the PILPS-Urban ones for the sensible and latent heat flux. Some models representing two-dimensional canyons now perform nearly as well as one and two-tile models after efforts to improve hydrology and vegetation representation. However, models with complex urban geometry often still have relatively simple hydrology and vegetation and perform less well overall (Lipson et al., 2023a). Suggesting the representation of hydrology and vegetation requires more attention (Lipson et al., 2023a).

Although PILPS-Urban and Urban-PLUMBER conclude vegetation and hydrology are important for model performance, neither project evaluates the water balance explicitly. The water balance satisfies the conservation of mass (Lavoisier, 1789) in the same way the energy balance satisfies the conservation of energy (Châtelet, 1740). The conservation of energy is forced in many ULSMs to prevent the energetic state of the model from drifting and the consequential, long-term bias in the modeled surface fluxes (Grimmond et al., 2010). Closure is achieved by either updating the surface temperatures based on the residual energy or restricting the turbulent heat flows to the available energy (Grimmond et al., 2010). Both PILPS-Urban and Urban-PLUMBER test whether models close the energy balance, but have not verified the numerical closure of the water balance. Similar to the energy balance, an unclosed water balance can result in model biases and consequential drifting. These biases may in turn affect the energy balance, as the energy and water balance are linked through evapotranspiration ($ET$), the mass counterpart of the latent heat flux ($Q_E$). This direct link implies errors and/or biases in one balance will affect the model's skill for the other balance. Recently, Yu et al. (2022) showed the hydrology in a coupled ULSM has the potential to improve the $Q_E$, humidity, and air temperature with impacts up into the boundary layer ($\sim$1 km). $ET/Q_E$ has been amongst the most challenging fluxes for ULSMs from the first assessment (Ross & Oke, 1988) until now (Grimmond et al., 2011). Given the link to the energy balance, closing the water balance may improve model performance for the energy balance fluxes.

However, the water balance cannot be directly assessed because of a lack of observations at the appropriate spatiotemporal scales at this time. While precipitation is measured routinely in many urban locations with rain gauges and rain radars, runoff, irrigation, and changes in water storage are not. $Q_E$ ($ET$) observations from eddy-covariance systems have substantial gaps introduced in the quality control process (Feigenwinter et al., 2012) that rejects more data close to rain events (Grimmond, 2006). Runoff is occasionally measured in urban catchments (Berthier et al., 1999; Walsh et al., 2005), but a challenge is posed by the difference in the source area of observations for runoff and eddy-covariance techniques (Grimmond & Oke, 1986, 1991; Hellsten et al., 2015). External water use, often irrigation, further complicates the water balance in cities, as it mainly occurs at the micro-scale (e.g. garden irrigation). This scale can only be inferred from neighborhood piped water supply observations and water use surveys or estimated from weather, vegetation, and soil type (Grimmond & Oke, 1986; Mitchell et al., 2001; Zeisl et al., 2018; Kokkonen et al., 2018). Tree roots penetrate (sewer) pipes causing damage (Randrup et al., 2001) and simultaneously taking out water, which is an unobserved term. Lastly, measuring the water storage change is logistically difficult, as this requires the state of each individual element contributing to water storage in the city, such as soil moisture, interception, groundwater, and surface water. Thus, a direct comparison of a full set of water balance observations is extremely challenging and an alternative approach is needed.

Here, we develop an alternative approach to evaluate the representation and dynamics of the water balance in ULSMs. To examine the water balance closure, we propose an UWBR (urban water balance representation) score. The score combines seven indicators assessing: water balance closure (1 indicator), $ET$ (2), water storage dynamics (2), and surface runoff (2). The UWBR score is applied, given a lack of observations, to rank models' capability to accurately capture different aspects of the water balance. Assessing the score of 19 Urban-PLUMBER ULSMs with a complete water balance representation helps to identify model improvement possibilities. The water balance representation is compared with the turbulent heat fluxes model skill since we expect a better water balance representation should improve simulated latent heat fluxes.

# 2 Methods

## 2.1 Urban water balance representation (UWBR) score

The UWBR score is a linear sum of seven indicators of a good water balance, which are assigned a value of 1 if a specified threshold is passed (Table 1), except the $I_{S,m}$ indicator, for which both sub-metrics are assigned 0.5 if passed. No weights are assigned, as these cannot be determined objectively. The UWBR score is compared with the model performance for the latent heat flux assessed with metrics capturing different characteristics (Willmott, 1982) that are not entirely independent:

- Absolute mean bias error (|MBE|) assesses the bias providing insight into how well the quantities of the latent heat flux are modeled.
- Coefficient of determination ($R^2$) captures the consistency of the timing as $R^2$ decreases with a shift in a quasiperiodic signal like the latent heat flux.
- Normalized standard deviation ($\sigma_{norm}$, $\sigma_{model}$ divided by $\sigma_{observations}$) compares the variability, which is dominated by the daily cycle in the case of the latent heat flux.
- Systematic Mean Absolute Error ($MAE_s$) indicates the average error. The systematic error is separated from the unsystematic error similarly to the approach presented by Willmott (1982) for the root mean square error. This separation allows us to distinguish between systematic and random errors.
- Unsystematic Mean Absolute Error ($MAE_u$) assesses how well the erratic behaviour is captured.

Before the individual indicators are introduced, we define two ways to calculate water storage from the model output based on either the water storage term or the other terms of the water balance combined. Assuming that the net change in water stored in a "catchment" or a model grid ($\Delta S$) can be derived from the difference between the incoming and outgoing water fluxes, then:

$$\Delta S = P + I - (R + ET) \quad (1)$$

where $P$ is precipitation, $I$ irrigation, and $R$ runoff. $R$ represents both the surface ($R_s$) and the subsurface ($R_{sub}$) runoff. When $\Delta S$ is calculated from the fluxes on the right-hand side of Eq. 1, we refer to this as the net water flux. Following the urban water balance (Grimmond & Oke, 1986), the net storage change ($\Delta S$) should account for the water storage change above and below ground, such as the interception, water bodies, and groundwater. The components actually included depend on the model conceptualization. Here, we refer to the storage represented in the model as the water storage ($\Delta S_{model}$):

$$\Delta S_{model} = \Delta S_{soil} + \Delta S_{intercept} + \Delta S_{snow} \tag{2}$$

where $\Delta S_{soil}$ is storage change in the soil moisture, $\Delta S_{intercept}$ storage change in the interception storage, and $\Delta S_{snow}$ storage change in the snow cover. When we refer to annual timescales, the analysis is performed on all time intervals of a year in the time series, i.e. a new annual period starts at every timestep, after which a full year of data is available (e.g. NL-Amsterdam: 2018-05-01 19:00 - 2019-05-01 19:00, 2018-05-01 20:00 - 2019-05-01 20:00, etc.). This method maximizes the use of available data and eliminates the influence of choosing a specific annual period like the calendar or hydrological year.

#### 2.1.1 Water balance closure

Water balance closure assumes that all fluxes add up to zero for the time and space under consideration (here $\sim 1$ km$^2$ and one year):

$$P + I - (R + ET + \Delta S) = 0 \tag{3}$$

where $\Delta S$ corresponds to the water storage in the model (Eq. 2) to prevent closure resulting from calculating the storage change based on the fluxes. Three models (8, 16, and 17) model groundwater interaction, which is not included in the model output. We examine the annual water balance closure with the annual total fluxes normalized by annual precipitation plus irrigation to enable comparison between sites with a range of precipitation regimes.

The water balance closure indicator ($I_A$, Table 1) assesses if the total sum of all fluxes (including storage) is less than 3% from P + I. The 3% threshold allows for non-closure due to unsaved interception storage data not being provided in the model output, errors arising in latent heat flux unit conversion, or numerical model errors. Interception storage is represented in all 19 models analyzed here, but only three model outputs provided the values. According to the literature, this may explain a non-closure of up to 0.5%(Klaassen et al., 1998; Wouters et al., 2015; Carlyle-Moses et al., 2020). Conversion of latent heat flux to $ET$ can vary by up to 2% depending on temperature and snow effects (Bringfelt, 1986; Petrucci et al., 2010). Not all models correct for these effects. To account for numerical model errors arising from discretization and time stepping (MacKay et al., 2022), we allow deviations of up to 0.5%.

#### 2.1.2 Evapotranspiration ($ET$)

The two $ET$ indicators address the magnitude and timing. Given gaps in $ET$ observations prevent direct comparison of total modeled $ET$ ($ET_{model}$) over a model period, we use one of the Lipson et al. (2023a) benchmark models. This allows a total $ET$ to be obtained without gaps. The Lipson et al. (2023a) benchmark model ($ET_{bench}$) is derived using multivariate ordinary least squares regressions with a K-means clustering approach. The K-means clustering approach is trained in-sample using 81 clusters on four variables: incoming shortwave radiation, air temperature, relative humidity, and wind speed (KM4-IS-SWdown-Tair-RH-Wind in Lipson et al., 2023a). To reduce the hourly MBE, wind speed is omitted at both Helsinki sites. At all sites, the MBE is below 1 $W\ m^{-2}$ and at most sites below 0.1 $W\ m^{-2}$ evaluated against available data.

**Table 1.** Overview of the seven indicators that are linearly combined in the UWBR score, which is used to evaluate the urban water balance representation in ULSMs. The criterion indicates what needs to be achieved to assign a value of 1 to the indicator or 0.5 per test in the case of $I_{S,m}$. The uncertainty criteria (*) are discussed in sections 2.1.2 and 2.1.4. The notation in the equations is defined in the corresponding subsections of section 2.1. The details on all indicators can be found in section 2.1.

| **Water balance flux** | **Indicator** | **Description** | **Criterion** | **Equation** |
|---|---|---|---|---|
| All | $I_A$ | Closure of the annual water balance assessed relative to the precipitation plus irrigation | $< 0.03$ | $\left\lvert \frac{P+I-(R+ET+\Delta S)}{P+I} \right\rvert$ |
| ET | $I_{ET,m}$ | Modeled cumulative $ET$ normalized by the benchmark $ET$ ($ET_{bench}$) over the whole model period | Within benchmark uncertainty* | $\frac{ET_{model}}{ET_{bench}}$ |
| | $I_{ET,t}$ | Similarity of $ET$ recession timescale distribution between model and observations from the whole model run | $p < 0.05$ | Kolmogorov-Smirnov test (Chakravarti et al., 1967) |
| $\Delta S$ | $I_{S,m}$ | Range over the whole model run in stored water derived from the modeled water storage and the net water flux compared to water storage capacity | < (50% of soil volume + 3 $mm$ interception) | Range in cumulative $\Delta S_{model}$ (Eq. 2) and $\Delta S$ (Eq. 1) |
| | $I_{S,t}$ | Coefficient of determination ($R^2$) between changes in modeled water storage and the net water flux over the whole model period | >0.9 | $R^2$ of changes in $\Delta S_{model}$ (Eq. 2) and changes in $\Delta S$ (Eq. 1) |
| $R_s$ | $I_{R,m}$ | Curve number ($CN$) from modeled runoff events and from site characteristics | Within $CN$ uncertainty* | $CN$ method (section 2.1.4) |
| | $I_{R,t}$ | Mean lag (hours) between center of mass from precipitation and surface runoff of all events | < 1 hour | $R_{s,centroid} - P_{centroid}$ |

Therefore, $ET_{bench}$ is assumed to provide a reasonable estimate of the total $ET$ flux over the model run for the $I_{ET,m}$ indicator (Table 1). We compare in $Q_E$ units rather than $ET$, eliminating unit conversions and calculate the cumulative $ET$ flux uncertainty from the benchmark based on (1) the benchmark MBE multiplied by the run duration, and (2) lack of energy balance closure associated with eddy-covariance observations (Franssen et al., 2010; Foken et al., 2012; Mauder et al., 2020). The lack of energy closure is calculated by the net all-wave radiation minus the sum of the turbulent heat fluxes. If a lack of closure occurs, the unexplained energy is split between $Q_E$ and $Q_h$ on the Bowen ratio (Twine et al., 2000; Hirschi et al., 2017; Mauder et al., 2020). The $Q_E$ share is combined with the MBE multiplied by the run duration to form the benchmark uncertainty yielding a maximum uncertainty, as some energy will go to the storage heat flux. A model run passes $I_{ET,m}$ when $ET_{model}$ falls within the uncertainty of $ET_{bench}$.

The timing of modeled $ET$ is assessed assuming exponential $ET$ recession after rainfall based on the recession timescale estimated following the Jongen et al. (2022) methodology. This methodology considers only the first ten days to exclude the influence of longer dry periods and irrigation. A daily-timescale analysis circumvents observational gaps. Model and observations are assessed if they have the same distribution for the recession timescale with a Kolmogorov-Smirnov test (Chakravarti et al., 1967). The $I_{ET,t}$ indicator is assigned a value of 1 when the p-value is below 0.05.

### *2.1.3 Water storage*

Indicator $I_{S,m}$ evaluates the water storage by comparing the modeled water storage and cumulative net water flux ranges (Section 2.1) over the analysis period with respect to the estimated water storage capacity. According to the literature, soil water storage capacity is maximally half the soil depth for all soil types (Saxton et al., 1986). As the modeled soil depth depends on the model run, the soil water storage capacity is calculated for each separately. To account for interception storage, 3 $mm$ is added to the estimated water storage capacity based on tree and impervious interception observations (Klaassen et al., 1998; Wouters et al., 2015; Carlyle-Moses et al., 2020). The two models not including soil moisture do not pass the first check of this indicator and are only evaluated based on the net water flux (Table 2). Other models receive 0.5 score when either the modeled water storage range or the net cumulative water flux range falls within the estimated water storage capacity (or 1 for both).

Indicator $I_{S,t}$ quantifies the internal temporal consistency between the change in water storage (Eq. 2) and the net water flux (Eq. 1), which should be indicating the same flux. The coefficient of determination $R^2$ (Willmott, 1982) is calculated using storage changes using 30-min (or 60-min) model output depending on the site forcing data. This metric equals 1 if the timing between two fluxes is similar ($R^2 > 0.9$) independent of the flux bias, unlike other indicators (e.g. $I_A$ ). The two models without soil moisture output are assigned a value of 0 for $I_{S,t}$ as their performance could not be evaluated.

### *2.1.4 Surface runoff ($R_s$)*

Indicator $I_{R,m}$ assesses the $R_s$ magnitude relating total event precipitation to $R_s$ (Figure 1a). Without runoff observations, curve numbers ($CN$) are derived to evaluate modeled total event $R_s$ (Cronshey et al., 1985) based on the relation between the total event precipitation ($P_e$) and the total event $R_s$ ($R_e$):

$$R_e = \frac{(P_e - 0.2S)^2}{P_e + 0.8S} \text{ with } S = \frac{1000}{CN} - 10 \tag{4}$$

where $S$ is the potential maximum retention. To determine when precipitation events are independent, the auto-correlation of precipitation events is examined. A dry period of five hours (Figure S1) is assumed across all sites, which is consistent with Wenzel Jr

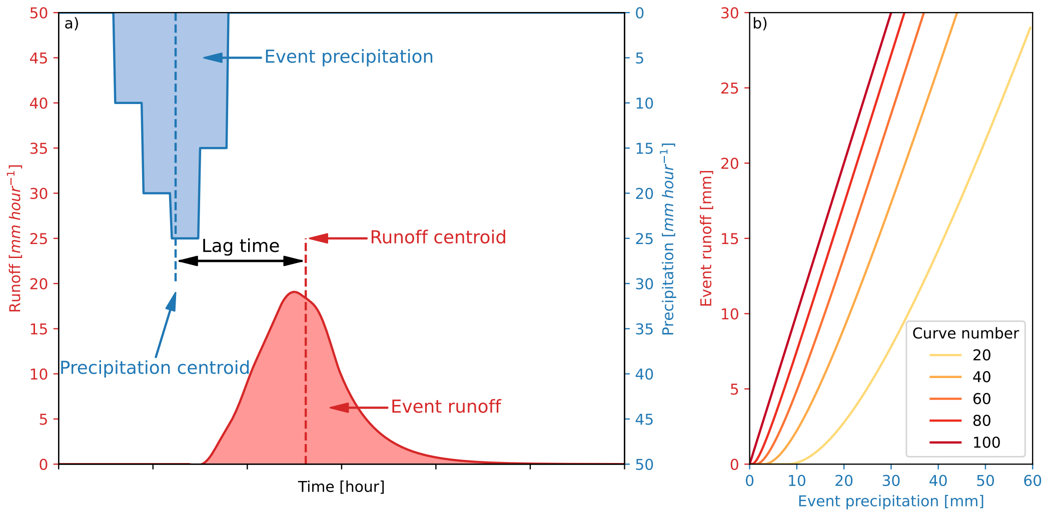


**Figure 1.** Illustration of surface runoff indicators ($I_{R,m}$ and $I_{R,t}$) showing (a) lag time between an event precipitation centroid and surface runoff centroid, and (b) $CN$ values (Eq. 4) derived from total event precipitation and surface runoff.

and Voorhees (1981). To exclude snow events, the analysis includes only events with a minimum air temperature above 0°C. For each model run, Eq. 4 is fit through the point cloud of $R_e$ versus $P_e$ to estimate $S$ and a standard deviation associated with the curve fit (Figure 1b). The $CN$ is derived from the $S$ estimate from the curve-fitting and the standard deviation is scaled accordingly to yield a $CN$ uncertainty estimate.

For each site, the $CN$ is estimated using a linear interpolation of a look-up table considering the impervious fraction within the eddy-covariance footprint (Cronshey et al., 1985). Given soil texture influences $CN$, sand fraction (Brakensiek & Rawls, 1983; Nachtergaele, 2001) obtained from a global data set (OpenLandMap, (Hengl, 2018)) is used to constrain $CN$ and provide uncertainty margins assuming an uncertainty of one-third of the $CN$ change in both directions from a one-level change in soil texture. If the site $CN$ uncertainty overlaps with the model $CN$ uncertainty, $I_{R,m}$ is assigned a value of 1.

Indicator $I_{R,t}$ addresses the rainfall-$R_s$ response times (Leopold, 1968). The lag time is calculated as the difference between centroids of rainfall ($P_{centroid}$) and $R_s$ ($R_{centroid}$) for the same events as the $CN$ calculations (Figure 1a). Long-tail rainfall events are excluded when the $R_{centroid}$ comes before the $P_{centroid}$. As eddy-covariance systems have a footprint on the sub-square-kilometer scale (Feigenwinter et al., 2012), lag time is expected to be much faster than 30-60 minutes (Morin et al., 2001; Berne et al., 2004; Yao et al., 2016), which is the model output resolution (Lipson et al., 2023a). Therefore, the mean lag time needs to be less than one hour. The mean is preferred over the median to also pinpoint models that occasionally have long lag times which would not affect the median.

### 2.2 Models

The present study anonymously analyzes the water balance outputs from 19 Urban-PLUMBER ULSMs (Table 2). Other Urban-PLUMBER ULSMs did not submit the necessary outputs to allow for a water balance assessment. The outputs are for 20 sites covering a range of climates, impervious fractions, and observational periods (Table 3). As two models did not run all sites, 377 runs are analyzed.

For each site, modelers were provided with the site characteristics and meteorological forcing with 10-year spin-up data (Lipson et al., 2022a). The spin-up period required to reach equilibrium varies per model, with some requiring many years to come to hydrological equilibrium with the forcing meteorology (Yang et al., 1995; Best & Grimmond, 2016). The 10 years of spin-up before the evaluation observations allowed the soil moisture stores to equilibrate with local conditions prior to analysis. ERA5 reanalysis data (Hersbach et al., 2020) are used to derive hourly forcing with bias-correction including diurnal and seasonal effects for each site (Lipson et al., 2022a).

Depending on site data, evaluation is undertaken with 30- or 60-minute fluxes for periods varying between 148 and 1827 days (average 912 days, Table 3). Similar to the Urban-PLUMBER protocol, to minimize human errors, modelers received a preliminary analysis of the water balance to help identify major issues and were encouraged to update their results. This eliminated unit errors, added missing variables, and removed inactive soil moisture layers.

For this study, we harmonize the hydrological model output. If a model only provided $Q_E$ (unit: $[W\ m^{-2}]$), it is converted to $ET$ (unit: $[mm\ d^{-1}]$) using latent heat of vaporization accounting for air temperature (Bringfelt, 1986). When snow is present the latent heat of fusion is added to the latent heat of vaporization to acquire the latent heat of sublimation (Petrucci et al., 2010). In the forcing, precipitation is split into snowfall and rainfall. At only 30% of the sites, snowfall amounts to more than 10% of the precipitation. It is added as rainfall for one model without snow hydrology, while the two others do not account for this input. Irrigation is simulated in two models. For all other models, it is assumed to be zero.

## 3 Results

The 19 ULSMs show a wide spread in the average yearly water fluxes at all 20 sites based on all 377 model runs (Figure 2). Overall, the model spread (whiskers, Figure 2) is wider than the modeled ensemble mean flux (bars, Figure 2). Models show more variation in $ET$ than in runoff. Sites with higher annual water input have more variability in model output fluxes, for example, the relatively high fluxes in KR-Jungnang and SG-TelokKurau compared to the lower yearly fluxes in PL-Lipowa and US-WestPhoenix.

### 3.1 Water balance closure

Although the annual mean model ensemble almost closes the water balance at most sites (Figure 2), most individual models do not close the water balance (Figure 3). Here, closure is assumed when the sum of all fluxes (Eq. 3) is less than 3% of P+I. This occurs in 57% of the model runs ($I_A$, Figure 4). In 25% of the model runs, non-closure exceeds 10% of P+I. Closure is model-related as the bias is similar across sites for each model (Figure 3). Five models close the water balance in all runs, whereas four models account for 48% of unclosed model runs. To assess the impact of model run length, the analysis is repeated with sites with more than two years of observations yielding similar results.

### 3.2 Evapotranspiration (*ET*)

Comparison of the modeled average diurnal range of the $ET$ (Figure 5) shows the highest inter-model spread at the peak of the daily cycle, with a range of 10-600% of the model ensemble-mean flux. Along three sites with contrasting precipitation regimes (US-WestPhoenix, AU-Preston, and SG-TelokKurau), $ET$ increases as expected at wetter sites.

**Table 2.** Overview of the 19 urban land surface models in the water balance analysis based on Lipson et al. (2023a). Two models did not provide soil moisture output ([1]). Three models capable of simulating irrigation did not include it in their Urban-PLUMBER runs ([2]).

| **Model** | **Urban geometry** | **Vegetation** | **Soil hydrology** | **Snow accumulation** | **Irrigation** | **Reference** |
|---|---|---|---|---|---|---|
| ASLUMv2.0 | Canyon | Grass | Multi-layer | No | No[2] | Z.-H. Wang et al. (2013)<br>C. Wang et al. (2021) |
| ASLUMv3.1 | Canyon | Grass+trees | Multi-layer | No | No[2] | Z.-H. Wang et al. (2013)<br>C. Wang et al. (2021) |
| CABLE | Non-urban | Separate tiles | Multi-layer | Veg. | No | Kowalczyk et al. (2006)<br>Y. P. Wang et al. (2011) |
| CHTESSEL | Non-urban | Separate tiles | Multi-layer | Veg. | No | Balsamo et al. (2009)<br>Boussetta et al. (2013) |
| CHTESSEL-U | Two-tile | Separate tiles | Multi-layer | Veg.+urban | No | McNorton et al. (2021)<br>Balsamo et al. (2009) |
| CLMU5 | Canyon | Grass+shrubs | Multi-layer | Urban | No | Oleson and Feddema (2020) |
| JULES 1T | One-tile | Separate tiles | Multi-layer | Veg.+urban | No | Best et al. (2011) |
| JULES 2T | Two-tile | Separate tiles | Multi-layer | Veg.+urban | No | Best et al. (2011) |
| JULES MOR | Two-tile | Separate tiles | Multi-layer | Veg.+urban | No | Best et al. (2011) |
| Lodz-SUEB | One-tile | Lumped with urban | Multi-layer[1] | Veg.+urban | No | Fortuniak (2003) |
| Manabe 1T | One-tile | Manabe bucket | One-layer | Veg.+urban | No | Best et al. (2011)<br>Manabe (1969) |
| Manabe 2T | Two-tile | Manabe bucket | One-layer | Veg.+urban | No | Best et al. (2011)<br>Manabe (1969) |
| NOAH-SLAB | One-tile | Separate tiles | Multi-layer | Veg.+urban | No | Kusaka et al. (2001)<br>Ek et al. (2003) |
| NOAH-SLUCM | Canyon | Separate tiles | Multi-layer | Veg.+urban | No | Kusaka et al. (2001)<br>Ek et al. (2003) |
| SNUUCM | Canyon | Separate tiles | Multi-layer[1] | Veg. | No | Ryu et al. (2011)<br>Ek et al. (2003) |
| SUEWS | Two-tile | Separate tiles | One-layer | Veg.+urban | No[2] | Järvi et al. (2011)<br>Ward et al. (2016) |
| TERRA 4.11 | One-tile | Separate tiles | Multi-layer | Veg. | No | Wouters et al. (2015)<br>Schulz and Vogel (2020) |
| UCLEM | Canyon | Grass+shrubs | One-layer | Veg.+urban | Yes | Thatcher and Hurley (2012)<br>Lipson et al. (2018) |
| UT&C | Canyon | Grass+shrubs+trees | Multi-layer | No | Yes | Meili et al. (2020) |

**Table 3.** Model (Table 2) outputs are analyzed for 20 sites (Lipson et al., 2022a). Only urban wind directions are included for the Minneapolis site. Characteristics include the local climate zone (LCZ, Stewart and Oke (2012), where 2 is compact mid-rise, 3 compact low-rise, 5 open mid-rise, and 6 open low-rise), impervious surface fraction ($F_{imp}$), displacement height ($z_d$), and eddy-covariance sensor height above ground level ($z_s$).

| Country | City (site) | Name | Lat. (°) | Lon. (°) | Observed period (days) | Köppen-Geiger climate | LCZ | $F_{imp}$ | $z_d$ (m) | $z_s$ (m) | Reference |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Australia | Melbourne (Preston) | AU-Preston | -37.73 | 145.01 | 475 | Cfb | 6 | 0.62 | 8 | 40 | Coutts et al. (2007a)<br>Coutts et al. (2007b) |
| Australia | Melbourne (Surrey Hills) | AU-SurreyHills | -37.83 | 145.10 | 148 | Cfb | 6 | 0.54 | 8 | 38 | Coutts et al. (2007a)<br>Coutts et al. (2007b) |
| Canada | Vancouver (Sunset) | CA-Sunset | 49.23 | -123.08 | 1827 | Csb | 6 | 0.68 | 3 | 25 | Christen et al. (2011)<br>Crawford and Christen (2015) |
| Finland | Helsinki (Kumpula) | FI-Kumpula | 60.20 | 24.96 | 1096 | Dfb | mix | 0.46 | 6 | 31 | Karsisto et al. (2016) |
| Finland | Helsinki (Torni) | FI-Torni | 60.17 | 24.94 | 1096 | Dfb | 2 | 0.77 | 15 | 60 | Nordbo et al. (2013)<br>Järvi et al. (2018) |
| France | Toulouse (Capitole) | FR-Capitole | 43.60 | 1.45 | 375 | Cfa | 2 | 0.90 | 11 | 48 | Masson et al. (2008)<br>Goret et al. (2019) |
| Greece | Heraklion | GR-HECKOR | 35.34 | 25.13 | 367 | Csa | 3 | 0.92 | 17 | 27 | Stagakis et al. (2019) |
| Japan | Tokyo (Yoyogi) | JP-Yoyogi | 35.66 | 139.68 | 1461 | Cfa | 2 | 0.92 | 28 | 52 | Hirano et al. (2015)<br>Ishidoya et al. (2020) |
| South Korea | Seoul (Jungnang) | KR-Jungnang | 37.59 | 127.08 | 825 | Dwa | 3 | 0.97 | 15 | 42 | J.-W. Hong et al. (2020)<br>S.-O. Hong et al. (2023) |
| South Korea | Cheongju (Ochang) | KR-Ochang | 36.72 | 127.43 | 780 | Dwa | 5 | 0.47 | 4 | 19 | J.-W. Hong et al. (2019)<br>J.-W. Hong et al. (2020) |
| Mexico | Mexico City (Escandon) | MX-Escandon | 19.40 | -99.18 | 470 | Cwb | 2 | 0.94 | 8 | 37 | Velasco et al. (2011)<br>Velasco et al. (2014) |
| Netherlands | Amsterdam | NL-Amsterdam | 52.37 | 4.89 | 652 | Cfb | 2 | 0.68 | 10 | 40 | Steeneveld et al. (2020) |
| Poland | Łódź (Lipowa) | PL-Lipowa | 51.76 | 19.45 | 1827 | Dfb | 2 | 0.76 | 7 | 37 | Pawlak et al. (2011)<br>Fortuniak et al. (2013) |
| Poland | Łódź (Narutowicza) | PL-Narutowicza | 51.77 | 19.48 | 1827 | Dfb | 2 | 0.65 | 11 | 42 | Fortuniak et al. (2006)<br>Fortuniak et al. (2013) |
| Singapore | Singapore (Telok Kurau) | SG-TelokKurau | 1.31 | 103.91 | 366 | Af | 3 | 0.85 | 7 | 24 | Roth et al. (2017) |
| UK | London (King's college) | UK-KingsCollege | 51.51 | -0.12 | 638 | Cfb | 2 | 0.79 | 15 | 50 | Kotthaus and Grimmond (2014a)<br>Kotthaus and Grimmond (2014b)<br>Bjorkegren et al. (2015) |
| UK | Swindon | UK-Swindon | 51.58 | -1.80 | 715 | Cfb | 6 | 0.49 | 4 | 13 | Ward et al. (2013) |
| USA | Baltimore (Cub hill) | US-Baltimore | 39.41 | -76.52 | 1826 | Cfa | 6 | 0.31 | 4 | 37 | Crawford et al. (2011) |
| USA | Minneapolis | US-Minneapolis1 | 45.00 | -93.19 | 1093 | Dfa | 6 | 0.21 | 3 | 40 | Peters et al. (2011)<br>Menzer and McFadden (2017) |
| USA | Phoenix (West) | US-WestPhoenix | 33.48 | -112.14 | 382 | Bwh | 6 | 0.48 | 3 | 22 | Chow et al. (2014)<br>Chow (2017) |

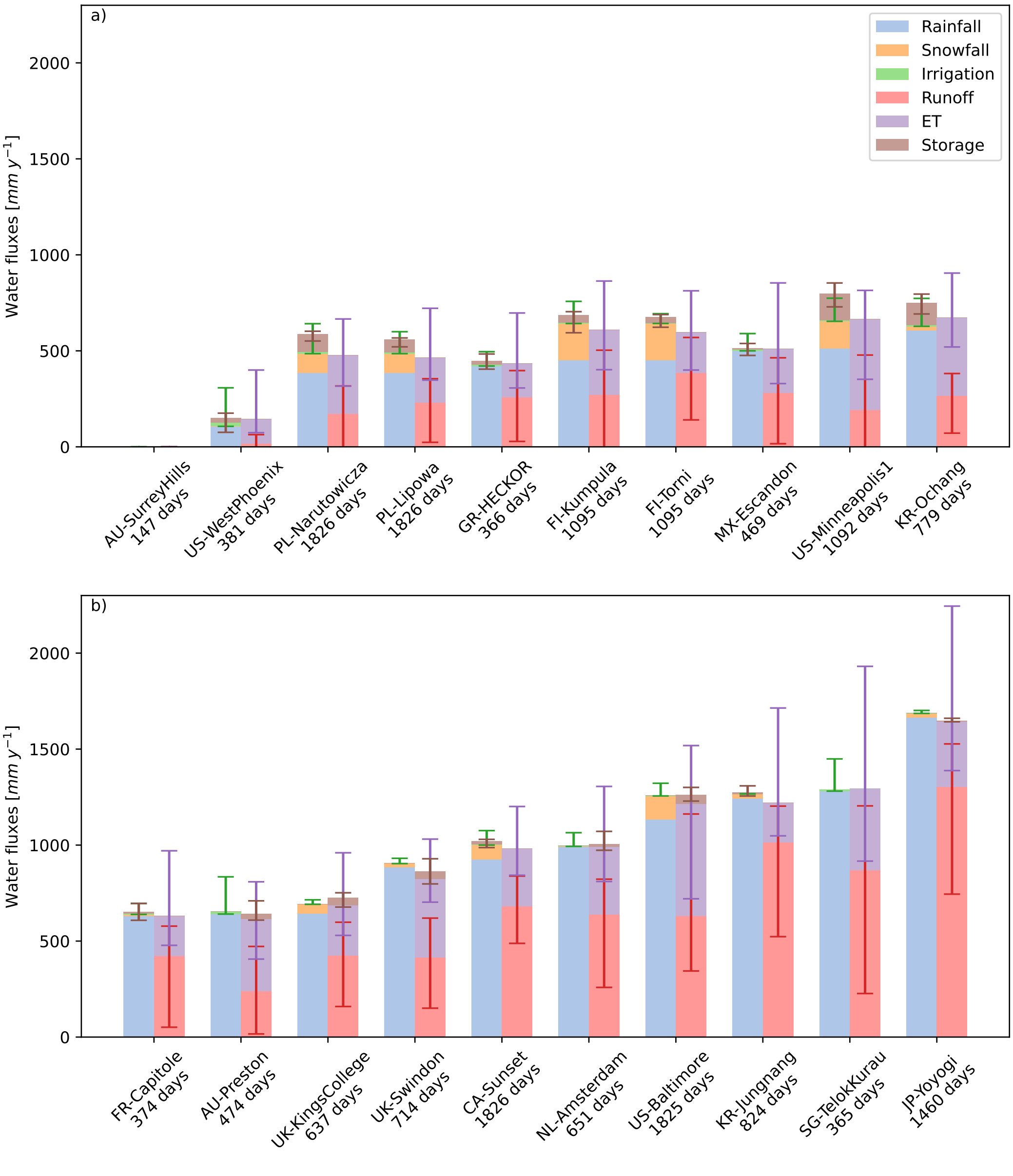


**Figure 2.** Ensemble mean (bars) and full range (minimum to maximum, whiskers) of the modeled annual water fluxes for all 20 sites ordered by increasing average annual precipitations. Modeled storage flux (Eq. 2, brown) appears on the left if a net input and right if a net loss. Values are means of all complete yeas in a data set (e.g. NL-Amsterdam: 2018-05-01 19:00 - 2019-05-01 19:00, 2018-05-01 20:00 - 2019-05-01 20:00, etc.). AU-Surreyhills has less than a year of observations.

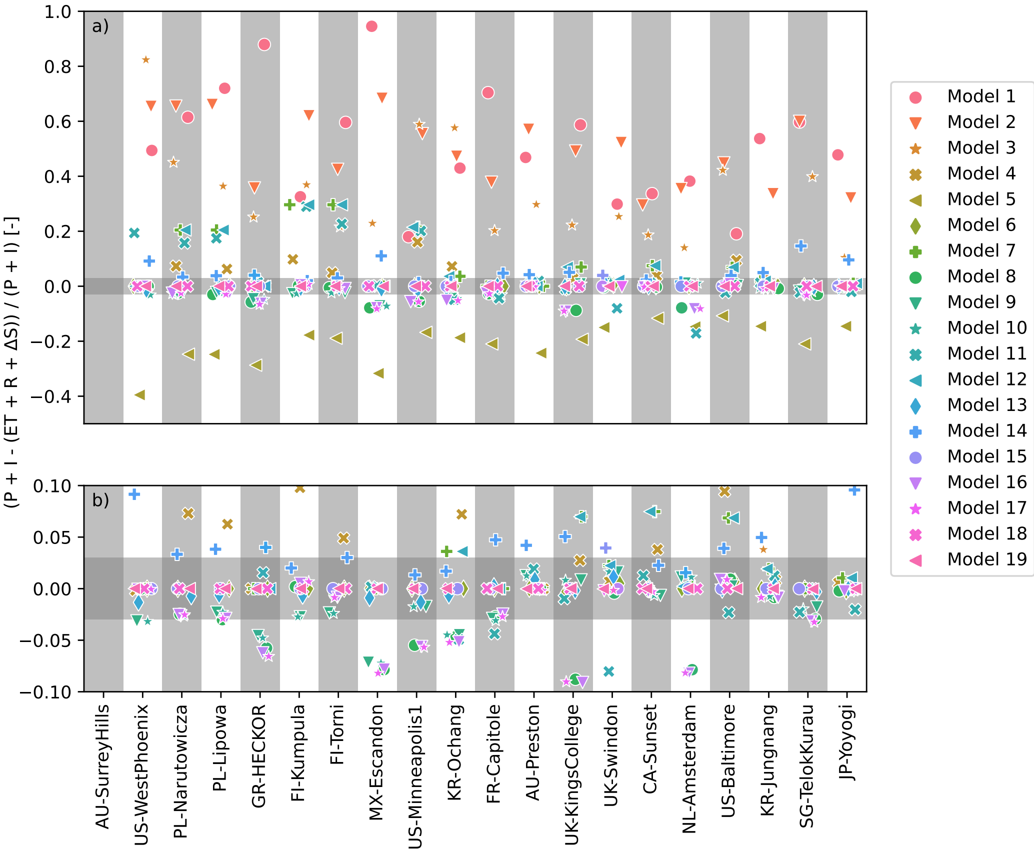


**Figure 3.** Annual water balance closure (Eq. 3) per model (marker) at 20 sites (by increasing average annual precipitation). Models with indicator $I_A = 1$ (Table 1, horizontal shading) are shown in more detail in the lower panel (b).

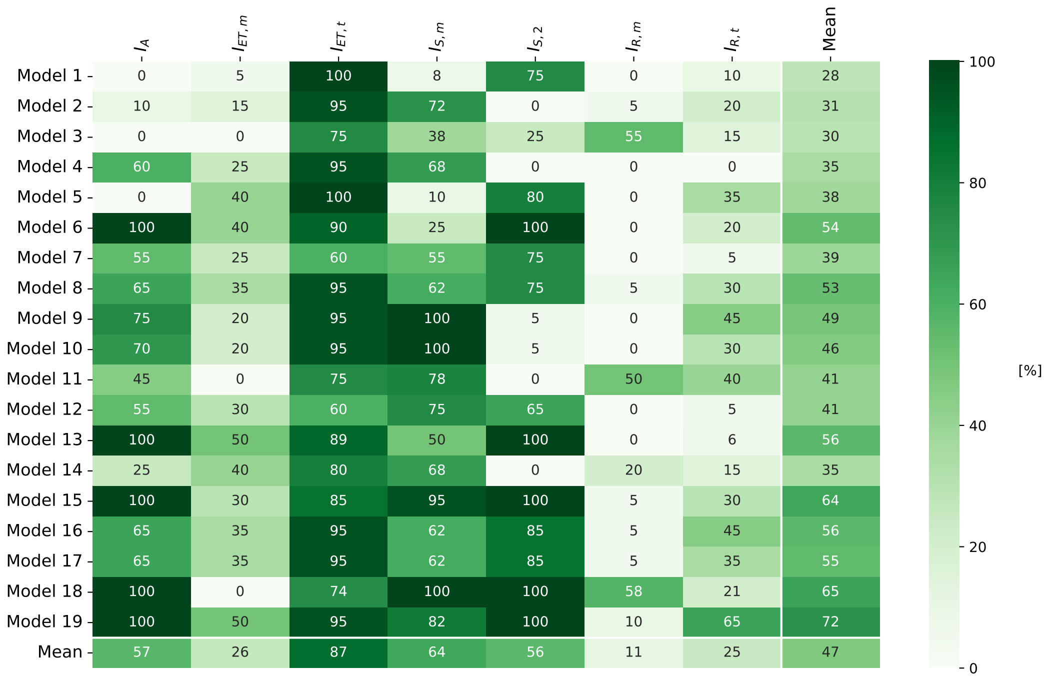


**Figure 4.** Overview of the indicators of the urban water balance representation (UWBR) score and constituent indicators (Table 1) over all sites. Means are corrected for missing model runs.

At US-WestPhoenix, all models but one underestimate $ET$. This underestimation likely results from the absence of irrigation in nearly all models, while irrigation is common at US-WestPhoenix (Templeton et al., 2018). At the other two sites, around half the models underestimate $ET$ (Figure 5). Although for these sites the model medians are better, the difficulty of capturing the correct flux magnitude is evident, as $I_{ET,m}$ is passed by only 26% of the model runs (Figure 4). No model passes this indicator at more than half of the sites.

After different rainfall events, daily $ET$ decreases with varying timescales in both the observations and the models (Figure 6). The variation is higher amongst the modeled than the observed drydowns. In contrast with the $ET$ magnitude, the recession timescale shows no link with the precipitation regime. $I_{ET,t}$ shows the $ET$ recession timescale is captured correctly in 87% of the cases (Figure 4).

### 3.3 Water storage

Not all models have water storage values (Eq. 2) that are equal to the cumulative net water flux (Eq. 1, Figure 7), which is seen across all sites (not shown). However, the water storage should reflect the cumulative net water flux, as the storage change is equal to the net water flux. For five models, the storage change is equal to the net water flux at all sites. Minor differences occur in six models and large differences in six others. Two models have no differences at sites without snowfall (e.g. AU-Preston) but large differences at sites with snowfall (e.g. CA-Sunset), as these models do not account for the snowfall in the input we see an increasing difference between the cumulative net water flux and the water storage. The models with larger differences follow a seasonal cycle likely caused by a non-restricted cumulative net water flux combined with restricted water storage by soil storage capacity.

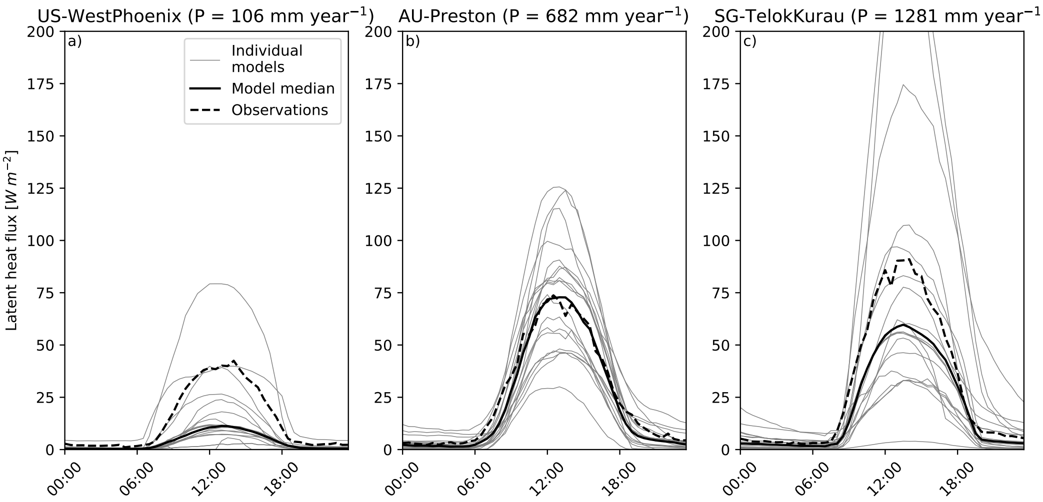


**Figure 5.** Illustration of modeled and observed (dashed) mean diurnal cycle of $ET$ at three sites with contrasting annual rainfall: (a) US-WestPhoenix, (b) AU-Preston, and (c) SG-TelokKurau. Note that the observations are direct latent heat flux observations from eddy-covariance systems and do not refer to $ET_{bench}$.

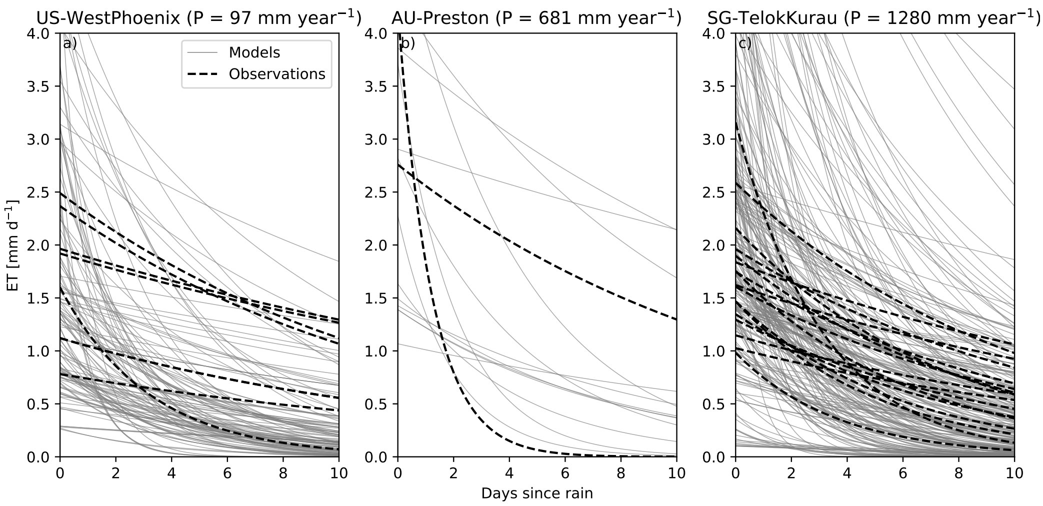


**Figure 6.** As Figure 5, but modeled (grey) and observed (black) daily $ET$ following separate, individual rainfall events. Drydown events are selected based on their duration and data availability (see Jongen et al., 2022). Note that the observations are direct latent heat flux observations from eddy-covariance systems and do not refer to $ET_{bench}$.

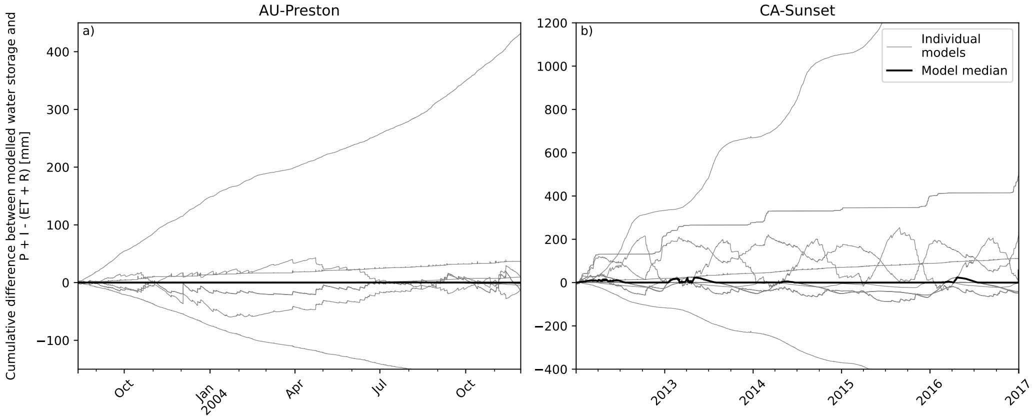


**Figure 7.** Cumulative difference between the water storage (Eq. 2) and cumulative net water flux (Eq. 1) at two representative sites for the entire model period for all models. Snowfall occurs at CA-Sunset, but not at AU-Preston. Some models are not visible as they are close to zero.

The range of modeled water storage exceeds the estimated site water storage capacity ($I_{S,m}$) in 64% of cases (Figure 4). Models 1 and 5 have the lowest score for this indicator, because they have an inconsistency between the inputs and outputs (Eq. 3) causing non-closure of the water balance at nearly all sites. Three models never exceed the estimated water storage capacity.

How water storage relates to cumulative net water flux is linked to the individual models given the consistent results across sites (Figure 9). With magnitude represented by water balance closure, we focus on the timing by assessing the water storage relative to the cumulative net water flux (Figure 8a-c). Model runs can have comparable directions but different patterns, e.g. model 11 (Figure 8a), comparable patterns but different magnitudes of change, e.g. model 9 (Figure 8b), or virtually no differences (e.g. model 18, Figure 8c). The water storage change and the net water flux (Figure 8d-f) emphasizes the differences in timing, which is why the indicator uses the $R^2$ of these derivatives. Only five models have virtually no differences and thus an $R^2$ of 1 (Figure 4). Over half of the models have $R^2$ greater than 0.9 indicating timing consistency ($I_{S,t}$, Figure 4).

### 3.4 Surface runoff ($R_s$)

All models have surface runoff triggered by precipitation, but the precipitation event size causing $R_s$ events differs between models (Figure 10). The model rather than the site seems to explain triggering event size despite the variation amongst sites in impervious fractions and precipitation regimes. This suggests that surface runoff parameterization may be critical. Thus, we find a large inter-model spread in the cumulative modeled $R_s$ (Figure 2). One model is excluded as it does not output $R_s$ separately from $R_{sub}$. Ten models show the expected increase of cumulative $R_s$ with increasing site impervious fraction (p>0.05, Wald test (Wald, 1943)), whereas nine models do not (Figure S2).

Only in 43 of the 337 model runs, the $CN$ (curve number: Section 2.1.4) is captured correctly, passing $I_{R,m}$ (Figure 4), so all other model runs have no overlap with the site estimates (see Section 2.1.4). Three models capture the $CN$ correctly for at least half of their model runs and are responsible for 32 of the successful model runs. Most models do not match event precipitation and $R_s$ relation. Most models underestimate the $CN$ relative to the site estimate (Figure S3). Underestimating the $CN$ indicates a

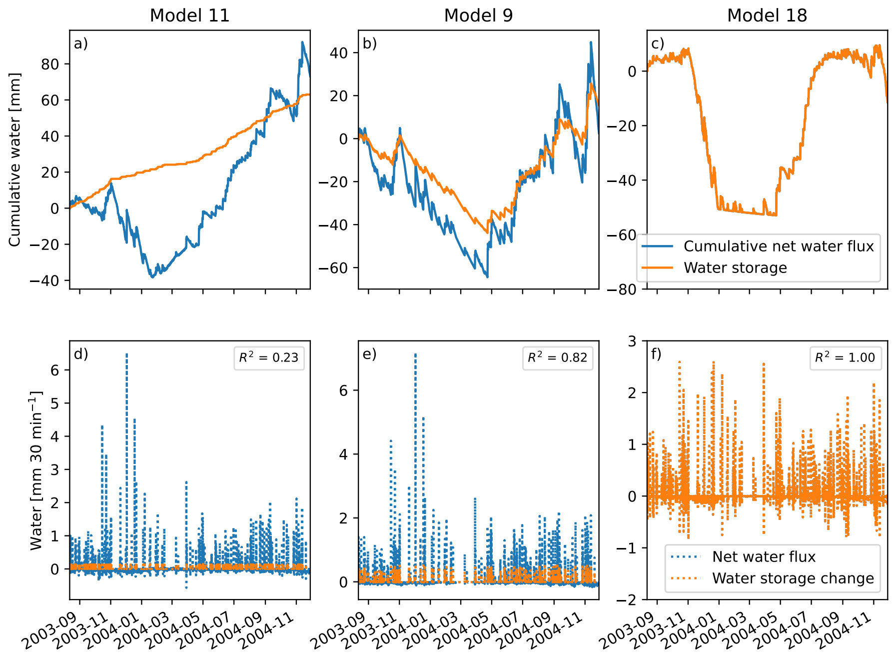


**Figure 8.** Illustration of the hourly water storage (Eq. 2) and the cumulative net water flux (Eq. 1) for 475 days at AU-Preston (a-c) and the water storage change and the net water flux (d-f) for three models with increasing coefficient of determination ($R^2$) of the water storage change and the net water flux determined at (half-)hourly resolution.

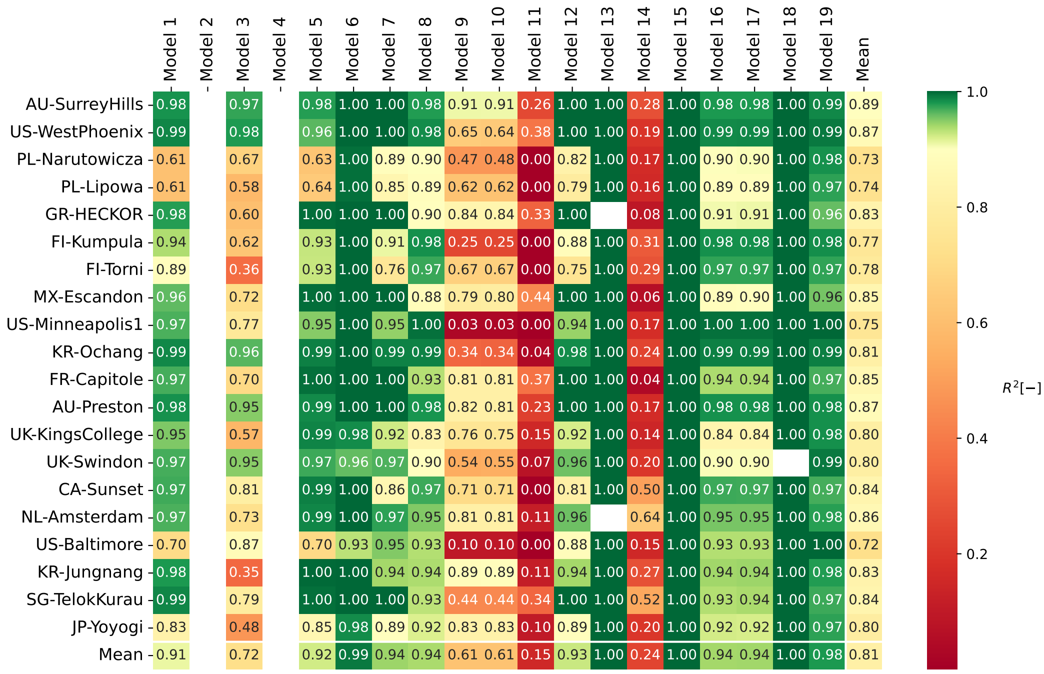


**Figure 9.** Coefficient of determination ($R^2$) between (half-)hourly water storage change (Eq. 2) and net water flux (Eq. 1) by model and site. Green indicates the 0.9 $I_{S,t}$ threshold (Table 1). Missing results are shown as white (i.e. cannot calculate water storage change or net water flux). Figure 8 may aid interpretation of $R^2$ values.

model is overestimating surface interception and/or soil infiltration, reducing $R_s$ (Equation 4).

One in four model runs accurately captures the fast $R_s$ response in the lag time (Figure 4) with $I_{R,t}$ passed by 25% of the model runs. With very short lag times expected, only overestimates are simulated. Most lag times averaged per model run are less than five hours, but exceptionally they are over 100 hours. Average lag times per model run are shown in Figure S4.

### 3.5 Urban water balance representation (UWBR) score

Across all model runs, the mean UWBR score amounts to 3.3 out of the possible 7 (Figure 4). Although the overall pass rate across all indciators and models is 47%, pass rates strongly vary per indicator. Notably, 87% passes $I_{ET,t}$, while only 11% passes $I_{R,m}$. Pass rates also differ among models from 28% to 72%. Only one model run passes all indicators, while 10 model runs have a score of 6 our of 7. Model 19 accounts for five of these eleven high-scoring runs. If a model closes the water balance ($I_A$), it generally scores better on both storage indicators. In contrast, models with a high passing percentage for one $ET$ indicator do not systematically score better for the other $ET$ indicator. Overall, the $ET$ timing ($I_{ET,t}$) is captured better than its cumulative magnitude ($I_{ET,m}$). A similar pattern is seen in the $R_s$ indicators with the timing ($I_{R,t}$) captured slightly better than magnitude ($I_{R,m}$).

Generally, pass rates per indicator show a dependence on the model (Figure 4). This dependence is not found for sites (Figure S5). There is no relation evident between UWBR score and model approach (e.g. built surface, soil hydrology, Table 2), but the model is

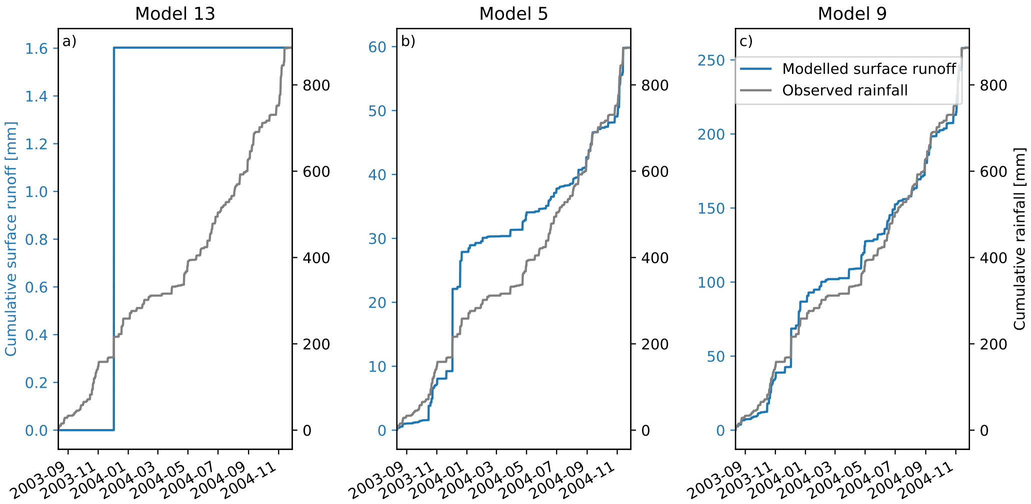


**Figure 10.** Illustration of surface runoff triggered for different AU-Preston precipitation events by three models (a) 13, (b) 5, and (c) 9. Note, the left-hand Y axis (surface runoff) increases (a→c), whereas the right-hand side Y axis (precipitation) is the same for all.

more influential than the site on UWBR score. As the Lipson et al. (2023a) classification (Table 2) was not developed with the water balance representation as its original goal, further work would be needed to identify what model attributes are key to better UWBR score.

### 3.6 Linking the water and energy balance

Surprisingly, models do not appear to capture any aspect of the latent heat flux more accurately if their UWBR score is higher. The UWBR score does not significantly correlate with better ranking on any of the four metrics evaluating the (half-)hourly modeled $Q_E$: the $R^2$, $\sigma_{norm}$, $RMSE_s$, and $RMSE_u$ (p>0.05, Wald test, Figure S6). These correlations remain absent if one of the indicators is omitted from the analysis. The lack of correlation may be the result of the low number (11) of runs with a UWBR score higher than 5 (Figure 4) effectively reducing the UWBR score range. Given the lack of relations between the UWBR score and $Q_E$ metrics, the $Q_E$ is not better captured in model runs that pass more indicators of a realistic water balance representation, thus refuting our hypothesis that the urban water balance skill positively impacts simulated energy fluxes.

## 4 Discussion and conclusions

This study assesses the water balance representation in 19 ULSMs from the Urban-PLUMBER project. It appears the water balance is not closed (within 3%) in 57% of the model-site runs. The considerable spread in water fluxes is as wide as the absolute flux magnitude at all sites. For both $ET$ and $R_s$, the timing is captured better than the flux magnitude. Modeled water storage dynamics (Eq. 2) are inconsistent with the net water flux (Eq. 1) in 44% of the models. Refuting our hypothesis, a better water balance representation does not result in more accurate latent heat fluxes. However, it is clear that the urban water balance is imperfectly incorporated into ULSMs and more proper physically-based representations are required.

Five models close the water balance at all sites (Models 6, 13, 15, 18, and 19), while three never reach closure (Models 1, 3, and 5). The other models close the water balance at some sites. For several non-closing models, we identify the causes. One model implicitly assumes an infinite source or sink of soil moisture by adapting the modeled soil moisture when it exceeds hard-coded limits adding or removing water to remain within these limits (Model 11). Two other models do not fully couple all processes, such as runoff and evaporation calculations occurring without water availability feedback between processes (Models 1 and 5). Such uncoupled processes may also explain inconsistent water storage dynamics and net water flux. Three models have groundwater flux, which is not included in the model output (Models 8, 16, and 17). One model without a snow module disregarded all snowfall creating a mismatch between real and modeled input (Model 2). For one model, we suspect a very shallow soil layer causes large numerical errors resulting in an unclosed water balance (Model 4). Fortunately, model improvements should be able to eliminate these issues for most models.

Evidence is found that the models would benefit from reevaluating their runoff parameterizations. The runoff volumes are poorly captured, resulting in $I_{R,m}$ having the poorest overall pass rate (Figure 4). Runoff has not been evaluated in previous ULSM comparisons and suffers here from a lack of direct observations and small areas being modeled (<1 km$^2$). The lack of correlation between modeled cumulative $R_s$ and the impervious fraction is worrying given the well-documented relation (Shuster et al., 2005; Jacobson, 2011). However, many models use relatively simple approaches, such as a constant fraction of rainfall that runs off independent of site characteristics, rainfall intensity, or soil moisture state. Others use poorly constrained parameters, such as how much water is routed between sub-grid tiles. Future work could help to constrain such parameters, while the simple approaches could be improved relatively straightforwardly.

Despite the lack of evidence showing a link between the UWBR score and $Q_E$ performance, the incomplete representation of the water balance may contribute to the poor latent heat flux performance of the ULSMs. The design of the UWBR score may not be successful in revealing an existing link between the UWBR score and $Q_E$ performance, as the UWBR score indicators assess the water balance based on physical realism and expectations derived from the literature. While a higher UWBR score indicates a more physically consistent water balance, it may still be an incorrect simulation. The opposite is also true, as, without physical constraints, machine learning approaches show good results for $Q_E$ (Vulova et al., 2021). Apart from that, a potential link between the water balance representation and the $Q_E$ performance may be hidden by other elements affecting $Q_E$ performance. These elements could be other components of the model (e.g. the energy balance representation) or human errors. Yet, we do find a poor performance for $Q_E$ consistent with the literature showing $Q_E$ is among the most challenging fluxes to model (Grimmond et al., 2011; Lipson et al., 2023a). As the energy and water balance are directly connected, we hypothesize potential errors in the water balance are causing, and not being caused by, the poor performance of $Q_E$, as the short runoff timescales in urban areas on a neighborhood scale dictate the water availability for $Q_E$ and not the other way around. Hence, good model performance for the latent and sensible heat flux cannot be achieved without properly representing both balances. Thus, we believe an improved representation of the water balance will assist in latent heat flux simulation and other energy fluxes.

This first systematic analysis of urban water balance modeling is an opportunistic study taking advantage of model outputs, model characterizations, and observations gathered for the Urban PLUMBER project (Lipson et al., 2023a, 2022a). The Urban-PLUMBER setup affects this study via (1) the diversity of model outputs linked to their range of modeling approaches, and (2) a lack of observations for all the water balance terms. Intentionally, a wide range of modeling approaches are analyzed with both default parameters and provided parameters implemented by modelers (Lipson et al., 2023a), impacting the model results and performance. For example, numerical discretization of soil layers can cause a flawed, reduced moisture drydown linked to irregular soil layer depths that enhance evaporation (MacKay et al., 2022). Ongoing land surface model developments to capture and link more processes increase both their scope and complexity, but the number of differing aspects complicates a systematic analysis aiming to attribute performance to certain aspects (Fisher & Koven, 2020; Blyth et al., 2021). To minimize human error, Urban-PLUMBER allowed resubmission of model outputs after web-based and manual checks. As these checks did not address the water balance, we provided an additional basic analysis of the water balance results to catch other human errors with encouragement to resubmit updated outputs. Unfortunately, resubmission reduces but does not eliminate human errors. All differences other than the water balance representation hinder the attribution of the model performance to the water balance concept as they explain the large variety in model performance amongst models that capture the water balance equally accurately. Ideally, these differences would be eliminated by developing a multi-model framework in the future (Sadegh et al., 2019) and characterizing model types based on water balance approaches. Such a characterization could allow for teasing out more detailed strengths and weaknesses of water balance representations.

Lack of observations (e.g. runoff, soil moisture) prevents direct assessment for many water balance terms. Hence, we develop a new alternative using quantitative indicators. Each indicator addresses a water balance process and checks whether it complies with physical limits, the model itself, or previous research. We refrain from weighting the indicators to minimize the score subjectivity and prevent one indicator from controlling the outcome. The systematic removal of one of the seven indicators allows us to confirm the UWBR score is not driven by one indicator.

Here, we show ULSMs produce a wide range of water balance results but often do not realistically represent important hydrological processes. Although our results are for offline ULSMs, we expect the identified issues will persist in a coupled setting on any scale (e.g., with mesoscale atmospheric models). ULSMs could be improved by ensuring they close the water balance and updating runoff parameterizations. Ideally, future energy-water–carbon studies will try to gather both a wider range of observations but also modeled processes. This will aid improvement of model processes and their feedbacks. However, the complexity of the urban landscape (e.g. different definitions between eddy covariance footprints, and runoff catchments) will require nested model runs and observations to ensure consistency of all. We recommend routine assessment of water balance closure in ULSM development phase applying the indicators of the UWBR score. In a broader context, both model evaluations and comparisons should extend beyond the target variables of the model to all processes that directly influence these variables. This will benefit the broader delivery of integrated urban services (WMO, 2019) and facilitate urban resilience across time scales.

## 5 Open research

All observation data from this study are openly available at Zenodo via https://doi.org/10.5281/zenodo.6590 (Lipson et al., 2022b). Model results and benchmarks (Lipson & Best, 2022) for AU-preston are archived at Zenodo. Model results for the other sites are visualized at https://urban-plumber.github.io/sites and will be published together with phase 2.

**Acknowledgments**

We acknowledge the Urban-PLUMBER project team and all observation and modeling participants providing the data set for this research. We would like to thank Judith Boekee, Andrew Frost, and Valentina Marchionni for the fruitful discussions. Harro Jongen acknowledges this research was supported by the WIMEK PhD grant 2020. Mathew Lipson acknowledges support from the Australian Research Council (ARC) Centre of Excellence for Climate System Science (grant CE110001028), National Computational Infrastructure (NCI) Australia and the Bureau of Meteorology, Australia. Gert-Jan Steeneveld acknowledges support from the Amsterdam Institute for Advanced Metropolitan Solutions (AMS Institute, project VIR16002) and the Netherlands Organisation for Scientific Research (NWO, Project 864.14.007). Sue Grimmond acknowledges support from ERC Urbisphere (grant 855055). Matthias Demuzere was supported by the ENLIGHT project, funded by the German Research Foundation (DFG) under grant no. 437467569. Ting Sun is supported by UKRI NERC Independent Research Fellowship (NE/P018637/2). Ruidong Li is supported by CSC scholarship. Keith Oleson's contribution is based upon work supported by the NSF National Center for Atmospheric Research, which is a major facility sponsored by the U.S. National Science Foundation under Cooperative Agreement No. 1852977. Chenghao Wang acknowledges support from the National Science Foundation (NSF) under grant No. OIA-2327435 and the National Oceanic and Atmospheric Administration (NOAA) under grant No. NA21OAR4590361.

## References

Balsamo, G., Beljaars, A., Scipal, K., Viterbo, P., van den Hurk, B., Hirschi, M., & Betts, A. K. (2009). A revised hydrology for the ECMWF model: Verification from field site to terrestrial water storage and impact in the integrated forecast system. *Journal of Hydrometeorology*, *10*(3), 623–643.

Berne, A., Delrieu, G., Creutin, J.-D., & Obled, C. (2004). Temporal and spatial resolution of rainfall measurements required for urban hydrology. *Journal of Hydrology*, *299*(3-4), 166–179.

Berthier, E., Andrieu, H., & Rodriguez, F. (1999). The Rezé urban catchments database. *Water Resources Research*, *35*(6), 1915–1919.

Best, M. J., Abramowitz, G., Johnson, H., Pitman, A., Balsamo, G., Boone, A., … others (2015). The plumbing of land surface models: benchmarking model performance. *Journal of Hydrometeorology*, *16*(3), 1425–1442.

Best, M. J., & Grimmond, C. S. B. (2016). Modeling the partitioning of turbulent fluxes at urban sites with varying vegetation cover. *Journal of Hydrometeorology*, *17*(10), 2537–2553.

Best, M. J., Pryor, M., Clark, D., Rooney, G., Essery, R., Ménard, C., … others (2011). The Joint UK Land Environment Simulator (JULES), model description–part 1: energy and water fluxes. *Geoscientific Model Development*, *4*(3), 677–699.

Bjorkegren, A., Grimmond, C. S. B., Kotthaus, S., & Malamud, B. (2015). $CO_2$ emission estimation in the urban environment: Measurement of the $CO_2$ storage term. *Atmospheric Environment*, *122*, 775–790.

Blyth, E. M., Arora, V. K., Clark, D. B., Dadson, S. J., De Kauwe, M. G., Lawrence, D. M., … others (2021). Advances in land surface modelling. *Current Climate Change Reports*, *7*(2), 45–71.

Botzen, W., Martinius, M., Bröde, P., Folkerts, M., Ignjacevic, P., Estrada, F., … Daanen, H. (2020). Economic valuation of climate change–induced mortality: age dependent cold and heat mortality in the Netherlands. *Climatic Change*, *162*(2), 545–562.

Boussetta, S., Balsamo, G., Beljaars, A., Panareda, A.-A., Calvet, J.-C., Jacobs, C., … others (2013). Natural land carbon dioxide exchanges in the ECMWF integrated forecasting system: Implementation and offline validation. *Journal of Geophysical Research: Atmospheres*, *118*(12), 5923–5946.

Brakensiek, D. L., & Rawls, W. J. (1983). Green-Ampt infiltration model parameters for hydrologic classification of soils. In *Advances in irrigation and drainage: surviving external pressures* (pp. 226–233).

Bringfelt, B. (1986). *Test of a forest evapotranspiration model.* SMHI.

Carlyle-Moses, D. E., Livesley, S., Baptista, M. D., Thom, J., & Szota, C. (2020). Urban trees as green infrastructure for stormwater mitigation and use. *Forest-Water Interactions*, 397–432.

Chakravarti, I., Laha, R., & Roy, J. (1967). *Handbook of methods of applied statistics. volume i: Techniques of computation, descriptive methods, and statistical inference.* John Wiley & Sons, Incorporated.

Châtelet, E. (1740). *Institutions de physique.* Paris.

Chow, W. T. (2017). *Eddy covariance data measured at the CAP LTER flux tower located in the west Phoenix, AZ neighborhood of Maryvale from 2011-12-16 through 2012-12-31.* Environmental Data Initiative.

Chow, W. T., Volo, T. J., Vivoni, E. R., Jenerette, G. D., & Ruddell, B. L. (2014). Seasonal dynamics of a suburban energy balance in Phoenix, Arizona. *International Journal of Climatology*, *34*(15), 3863–3880.

Christen, A., Coops, N., Crawford, B., Kellett, R., Liss, K., Olchovski, I., ... Voogt, J. (2011). Validation of modeled carbon-dioxide emissions from an urban neighborhood with direct eddy-covariance measurements. *Atmospheric Environment*, *45*(33), 6057–6069.

Coutts, A. M., Beringer, J., & Tapper, N. J. (2007a). Characteristics influencing the variability of urban $CO_2$ fluxes in Melbourne, Australia. *Atmospheric Environment*, *41*(1), 51–62.

Coutts, A. M., Beringer, J., & Tapper, N. J. (2007b). Impact of increasing urban density on local climate: Spatial and temporal variations in the surface energy balance in Melbourne, Australia. *Journal of Applied Meteorology and Climatology*, *46*(4), 477–493.

Crawford, B., & Christen, A. (2015). Spatial source attribution of measured urban eddy covariance co 2 fluxes. *Theoretical and Applied Climatology*, *119*, 733–755.

Crawford, B., Grimmond, C., & Christen, A. (2011). Five years of carbon dioxide fluxes measurements in a highly vegetated suburban area. *Atmospheric Environment*, *45*(4), 896–905.

Cronshey, R., Roberts, R., & Miller, N. (1985). Urban hydrology for small watersheds (tr-55 rev.). In *Hydraulics and hydrology in the small computer age* (pp. 1268–1273).

Demuzere, M., Kittner, J., Martilli, A., Mills, G., Moede, C., Stewart, I. D., ... Bechtel, B. (2022). A global map of local climate zones to support earth system modelling and urban scale environmental science. *Earth System Science Data Discussions*, *2022*, 1–57.

Ek, M., Mitchell, K., Lin, Y., Rogers, E., Grunmann, P., Koren, V., ... Tarpley, J. (2003). Implementation of noah land surface model advances in the national centers for environmental prediction operational mesoscale eta model. *Journal of Geophysical Research: Atmospheres*, *108*(D22).

Feigenwinter, C., Vogt, R., & Christen, A. (2012). Eddy covariance measurements over urban areas. *Eddy covariance: A practical guide to measurement and data analysis*, 377–397.

Fisher, R. A., & Koven, C. D. (2020). Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems. *Journal of Advances in Modeling Earth Systems*, *12*(4), e2018MS001453.

Fletcher, T. D., Andrieu, H., & Hamel, P. (2013). Understanding, management and modelling of urban hydrology and its consequences for receiving waters: A state of the art. *Advances in water resources*, *51*, 261–279.

Foken, T., Leuning, R., Oncley, S. R., Mauder, M., & Aubinet, M. (2012). Corrections and data quality control. In *Eddy covariance* (pp. 85–131). Springer.

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., Kłysik, K., & Siedlecki, M. (2006). New measurements of the energy balance components in Łódź. In *Preprints, sixth international conference on urban climate, Göteborg, Sweden* (pp. 12–16).

Fortuniak, K., Pawlak, W., & Siedlecki, M. (2013). Integral turbulence statistics over a central European city centre. *Boundary-Layer Meteorology*, *146*, 257–276.

Franssen, H. H., Stöckli, R., Lehner, I., Rotenberg, E., & Seneviratne, S. I. (2010). Energy balance closure of eddy-covariance data: A multisite analysis for European FLUXNET stations. *Agricultural and Forest Meteorology*, *150*(12), 1553–1567.

Gasparrini, A., Guo, Y., Sera, F., Vicedo-Cabrera, A. M., Huber, V., Tong, S., ... others (2017). Projections of temperature-related excess mortality under climate change scenarios. *The Lancet Planetary Health*, *1*(9), e360–e367.

Goret, M., Masson, V., Schoetter, R., & Moine, M.-P. (2019). Inclusion of $CO_2$ flux modelling in an urban canopy layer model and an evaluation over an old European city centre. *Atmospheric Environment: X*, *3*, 100042.

Grimmond, C. S. B. (2006). Progress in measuring and observing the urban atmosphere. *Theoretical and Applied Climatology*, *84*(1), 3–22.

Grimmond, C. S. B., Best, M. J., Barlow, J., Arnfield, A., Baik, J.-J., Baklanov, A., ... others (2009). Urban surface energy balance models: model characteristics and methodology for a comparison study. In *Meteorological and air quality models for urban areas* (pp. 97–123). Springer.

Grimmond, C. S. B., Blackett, M., Best, M. J., Baik, J.-J., Belcher, S., Beringer, J., ... others (2011). Initial results from phase 2 of the international urban energy balance model comparison. *International Journal of Climatology*, *31*(2), 244–272.

Grimmond, C. S. B., Blackett, M., Best, M. J., Barlow, J., Baik, J., Belcher, S., ... others (2010). The international urban energy balance models comparison project: first results from phase 1. *Journal of applied meteorology and climatology*, *49*(6), 1268–1292.

Grimmond, C. S. B., & Oke, T. R. (1986). Urban water balance: 2. results from a suburb of Vancouver, British Columbia. *Water Resources Research*, *22*(10), 1404–1412.

Grimmond, C. S. B., & Oke, T. R. (1991). An evapotranspiration-interception model for urban areas. *Water Resources Research*, *27*(7), 1739–1755.

Grimmond, C. S. B., & Oke, T. R. (2002). Turbulent heat fluxes in urban areas: Observations and a local-scale urban meteorological parameterization scheme (lumps). *Journal of Applied Meteorology and Climatology*, *41*(7), 792–810.

Hamdi, R., Kusaka, H., Doan, Q.-V., Cai, P., He, H., Luo, G., ... others (2020). The state-of-the-art of urban climate change modeling and observations. *Earth Systems and Environment*, 1–16.

Hamdi, R., & Schayes, G. (2007). Validation of Martilli's urban boundary layer scheme with measurements from two mid-latitude European cities. *Atmospheric Chemistry and Physics*, *7*(17), 4513–4526.

Heaviside, C., Vardoulakis, S., & Cai, X.-M. (2016). Attribution of mortality to the urban heat island during heatwaves in the West Midlands, UK. *Environmental health*, *15*(1), 49–59.

Hellsten, A., Luukkonen, S.-M., Steinfeld, G., Kanani-Sühring, F., Markkanen, T., Järvi, L., ... Raasch, S. (2015). Footprint evaluation for flux and concentration measurements for an urban-like canopy with coupled lagrangian stochastic and large-eddy simulation models. *Boundary-Layer Meteorology*, *157*(2), 191–217.

Henderson-Sellers, A., McGuffie, K., & Pitman, A. (1996). The project for intercomparison of land-surface parametrization schemes (PILPS): 1992 to 1995. *Climate Dynamics*, *12*(12), 849–859.

Hengl, T. (2018). *Sand content in % (kg/kg) at 6 standard depths (0, 10, 30, 60, 100 and 200 cm) at 250 m resolution (version v0. 2) [data set].* Zenodo. doi: https://doi.org/10.5281/zenodo.2525662

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... others (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999–2049.

Hertwig, D., Grimmond, C. S. B., Hendry, M. A., Saunders, B., Wang, Z., Jeoffrion, M., ... others (2020). Urban signals in high-resolution weather and climate simulations: role of urban land-surface characterisation. *Theoretical and Applied Climatology*, *142*(1), 701–728.

Hirano, T., Sugawara, H., Murayama, S., & Kondo, H. (2015). Diurnal variation of $CO_2$ flux in an urban area of Tokyo. *Sola*, *11*, 100–103.

Hirschi, M., Michel, D., Lehner, I., & Seneviratne, S. I. (2017). A site-level comparison of lysimeter and eddy covariance flux measurements of evapotranspiration. *Hydrology and Earth System Sciences*, *21*(3), 1809–1825.

Hong, J.-W., Hong, J., Chun, J., Lee, Y. H., Chang, L.-S., Lee, J.-B., ... Joo, S. (2019). Comparative assessment of net $CO_2$ exchange across an urbanization gradient in Korea based on eddy covariance measurements. *Carbon Balance and Management*, *14*(1), 1–18.

Hong, J.-W., Lee, K., & Hong, J. (2020). *Observational data of Ochang and Jungnang in Korea, EAPL at Yonsei University [data set].*

Hong, S.-O., Kim, J., Byun, Y.-H., Hong, J., Hong, J.-W., Lee, K., ... Kim, Y.-H. (2023). Intra-urban variations of the $CO_2$ fluxes at the surface-atmosphere interface in the Seoul Metropolitan Area. *Asia-Pacific Journal of Atmospheric Sciences*, 1–15.

Howard, L. (1833). *The climate of london deduced from meteorological observations made in the metropolis and various places around it.* (Vol. 1). W. Phillips.

Ishidoya, S., Sugawara, H., Terao, Y., Kaneyasu, N., Aoki, N., Tsuboi, K., & Kondo, H. (2020). $O_2$: $CO_2$ exchange ratio for net turbulent flux observed in an urban area of Tokyo, Japan, and its application to an evaluation of anthropogenic $CO_2$ emissions. *Atmospheric Chemistry and Physics*, *20*(9), 5293–5308.

Jacobson, C. R. (2011). Identification and quantification of the hydrological impacts of imperviousness in urban catchments: A review. *Journal of Environmental Management*, *92*(6), 1438–1448.

Järvi, L., Grimmond, C. S. B., & Christen, A. (2011). The surface urban energy and water balance scheme (SUEWS): Evaluation in Los Angeles and Vancouver. *Journal of Hydrology*, *411*(3-4), 219–237.

Järvi, L., Rannik, Ü., Kokkonen, T. V., Kurppa, M., Karppinen, A., Kouznetsov, R. D., ... Wood, C. R. (2018). Uncertainty of eddy covariance flux measurements over an urban area based on two towers. *Atmospheric Measurement Techniques*, *11*(10), 5421–5438.

Jongen, H. J., Steeneveld, G.-J., Beringer, J., Christen, A., Chrysoulakis, N., Fortuniak, K., ... Teuling, R. (2022). Urban water storage capacity inferred from observed evapotranspiration recession. *Geophysical Research Letters*, *49*(3), e2021GL096069.

Karsisto, P., Fortelius, C., Demuzere, M., Grimmond, C. S. B., Oleson, K., Kouznetsov, R., ... Järvi, L. (2016). Seasonal surface urban energy balance and wintertime stability simulated using three land-surface models in the high-latitude city Helsinki. *Quarterly Journal of the Royal Meteorological Society*, *142*(694), 401–417.

Klaassen, W., Bosveld, F., & De Water, E. (1998). Water storage and evaporation as constituents of rainfall interception. *Journal of Hydrology*, *212*, 36–50.

Kokkonen, T., Grimmond, C. S. B., Christen, A., Oke, T., & Järvi, L. (2018). Changes to the water balance over a century of urban development in two neighborhoods: Vancouver, Canada. *Water Resources Research*, *54*(9), 6625–

6642.
Koopmans, S., Heusinkveld, B., & Steeneveld, G. (2020). A standardized physical equivalent temperature urban heat map at 1-m spatial resolution to facilitate climate stress tests in the Netherlands. *Building and Environment*, *181*, 106984.
Kotthaus, S., & Grimmond, C. S. B. (2014a). Energy exchange in a dense urban environment–part II: Impact of spatial heterogeneity of the surface. *Urban Climate*, *10*, 281–307.
Kotthaus, S., & Grimmond, C. S. B. (2014b). Energy exchange in a dense urban environment–part I: Temporal variability of long-term observations in central London. *Urban Climate*, *10*, 261–280.
Kowalczyk, E., Wang, Y., Law, R., Davies, H., McGregor, J., & Abramowitz, G. (2006). The CSIRO Atmosphere Biosphere Land Exchange (CABLE) model for use in climate models and as an offline model. *CSIRO Marine and Atmospheric Research Paper*, *13*, 42.
Krayenhoff, E. S., & Voogt, J. A. (2007). A microscale three-dimensional urban energy balance model for studying surface temperatures. *Boundary-Layer Meteorology*, *123*(3), 433–461.
Kusaka, H., Kondo, H., Kikegawa, Y., & Kimura, F. (2001). A simple single-layer urban canopy model for atmospheric models: Comparison with multi-layer and slab models. *Boundary-Layer Meteorology*, *101*(3), 329–358.
Lavoisier, A. L. (1789). *Traite elementaire de chimie*. Paris.
Lemonsu, A., Viguie, V., Daniel, M., & Masson, V. (2015). Vulnerability to heat waves: Impact of urban expansion scenarios on urban heat island and heat stress in Paris (France). *Urban Climate*, *14*, 586–605.
Leopold, L. B. (1968). *Hydrology for urban land planning: A guidebook on the hydrologic effects of urban land use* (Vol. 554). US Geological Survey.
Lipson, M. J., & Best, M. (2022). *Benchmarks for the urban-plumber model evaluation project phase 1 (au-preston) (versie v1) [data set]*. Zenodo. doi: https://doi.org/10.5281/zenodo.7330052
Lipson, M. J., Grimmond, C. S. B., Best, M., Abramowitz, G., Coutts, A., Tapper, N., … Pitman, A. J. (2023a). Evaluation of 30 urban land surface models in the Urban-PLUMBER project: Phase 1 results. *Quarterly Journal of the Royal Meteorological Society*.
Lipson, M. J., Grimmond, C. S. B., Best, M., Chow, W. T., Christen, A., Chrysoulakis, N., … others (2022b). *Site data archive for "harmonized gap-filled dataset from 20 urban flux tower sites" for the Urban-PLUMBER project (version v1) [data set]*. Zenodo. doi: https://doi.org/10.5281/zenodo.2525662
Lipson, M. J., Grimmond, C. S. B., Best, M. J., Chow, W. T., Christen, A., Chrysoulakis, N., … others (2022a). Harmonized gap-filled datasets from 20 urban flux tower sites. *Earth System Science Data Discussions*, 1–29.
Lipson, M. J., Grimmond, S., Best, M., Abramowitz, G., Coutts, A., Tapper, N., … Wang, Z.-H. (2023b). The Urban-PLUMBER model evaluation project: Phase 1 results. Sydney, Australia. ($11^{th}$ International Conference on Urban Climate)
Lipson, M. J., Thatcher, M., Hart, M. A., & Pitman, A. (2018). A building energy demand and urban land surface model. *Quarterly Journal of the Royal Meteorological Society*, *144*(714), 1572–1590.
MacKay, M. D., Meyer, G., & Melton, J. R. (2022). On the discretization of Richards equation in Canadian land surface models. *Atmosphere-Ocean*, 1–11.
Manabe, S. (1969). Climate and the ocean circulation: I. the atmospheric circulation and the hydrology of the earth's surface. *Monthly Weather Review*, *97*(11), 739–774.
Masson, V., Gomes, L., Pigeon, G., Liousse, C., Pont, V., Lagouarde, J.-P., … oth-

ers (2008). The canopy and aerosol particles interactions in Toulouse urban layer (CAPITOUL) experiment. *Meteorology and Atmospheric Physics*, *102*, 135–157.

Mauder, M., Foken, T., & Cuxart, J. (2020). Surface-energy-balance closure over land: a review. *Boundary-Layer Meteorology*, *177*(2), 395–426.

McNorton, J., Arduini, G., Bousserez, N., Agustí-Panareda, A., Balsamo, G., Boussetta, S., ... Hogan, R. (2021). An urban scheme for the ECMWF integrated forecasting system: Single-column and global offline application. *Journal of Advances in Modeling Earth Systems*, e2020MS002375.

Meili, N., Manoli, G., Burlando, P., Bou-Zeid, E., Chow, W. T., Coutts, A. M., ... others (2020). An urban ecohydrological model to quantify the effect of vegetation on urban climate and hydrology (UT&C v1. 0). *Geoscientific Model Development*, *13*(1), 335–362.

Menzer, O., & McFadden, J. P. (2017). Statistical partitioning of a three-year time series of direct urban net $CO_2$ flux measurements into biogenic and anthropogenic components. *Atmospheric Environment*, *170*, 319–333.

Mitchell, V. G., Mein, R. G., & McMahon, T. A. (2001). Modelling the urban water cycle. *Environmental Modelling & Software*, *16*(7), 615–629.

Morin, E., Enzel, Y., Shamir, U., & Garti, R. (2001). The characteristic time scale for basin hydrological response using radar data. *Journal of Hydrology*, *252*(1-4), 85–99.

Nachtergaele, F. (2001). Soil taxonomy—a basic system of soil classification for making and interpreting soil surveys. *Geoderma*, *99*(3-4), 336–337.

Nordbo, A., Järvi, L., Haapanala, S., Moilanen, J., & Vesala, T. (2013). Intra-city variation in urban morphology and turbulence structure in Helsinki, Finland. *Boundary-Layer Meteorology*, *146*, 469–496.

Oke, T. R. (1982). The energetic basis of the urban heat island. *Quarterly Journal of the Royal Meteorological Society*, *108*(455), 1–24.

Oleson, K. W., Bonan, G. B., Feddema, J., Vertenstein, M., & Grimmond, C. S. B. (2008). An urban parameterization for a global climate model. part I: Formulation and evaluation for two cities. *Journal of Applied Meteorology and Climatology*, *47*(4), 1038–1060.

Oleson, K. W., & Feddema, J. (2020). Parameterization and surface data improvements and new capabilities for the community land model urban (CLMU). *Journal of advances in modeling earth systems*, *12*(2), e2018MS001586.

Pawlak, W., Fortuniak, K., & Siedlecki, M. (2011). Carbon dioxide flux in the centre of Łódź, poland—analysis of a 2-year eddy covariance measurement data set. *International Journal of Climatology*, *31*(2), 232–243.

Peters, E. B., Hiller, R. V., & McFadden, J. P. (2011). Seasonal contributions of vegetation types to suburban evapotranspiration. *Journal of Geophysical Research: Biogeosciences*, *116*(G1).

Petrucci, R. H., Herring, F. G., & Madura, J. D. (2010). *General chemistry: principles and modern applications.* Pearson Prentice Hall.

Porson, A., Clark, P. A., Harman, I., Best, M. J., & Belcher, S. (2010). Implementation of a new urban energy budget scheme in the metum. part I: Description and idealized simulations. *Quarterly Journal of the Royal Meteorological Society*, *136*(651), 1514–1529.

Randrup, T. B., McPherson, E. G., & Costello, L. R. (2001). Tree root intrusion in sewer systems: A review of extent and costs. *Journal of Infrastructure Systems 7: 26-31*, *7*, 26–31.

Ronda, R., Steeneveld, G., Heusinkveld, B., Attema, J., & Holtslag, A. (2017). Urban finescale forecasting reveals weather conditions with unprecedented detail. *Bulletin of the American Meteorological Society*, *98*(12), 2675–2688.

Ross, S. L., & Oke, T. (1988). Tests of three urban energy balance models. *Boundary-Layer Meteorology*, *44*(1), 73–96.

Roth, M., Jansson, C., & Velasco, E. (2017). Multi-year energy balance and carbon dioxide fluxes over a residential neighbourhood in a tropical city. *International Journal of Climatology*, *37*(5), 2679–2698.

Ryu, Y.-H., Baik, J.-J., & Lee, S.-H. (2011). A new single-layer urban canopy model for use in mesoscale atmospheric models. *Journal of Applied Meteorology and Climatology*, *50*(9), 1773–1794.

Sadegh, M., AghaKouchak, A., Flores, A., Mallakpour, I., & Nikoo, M. R. (2019). A multi-model nonstationary rainfall-runoff modeling framework: analysis and toolbox. *Water Resources Management*, *33*(9), 3011–3024.

Saxton, K., Rawls, W., Romberger, J. S., & Papendick, R. (1986). Estimating generalized soil-water characteristics from texture. *Soil Science Society of America Journal*, *50*(4), 1031–1036.

Schulz, J.-P., & Vogel, G. (2020). Improving the processes in the land surface scheme TERRA: Bare soil evaporation and skin temperature. *Atmosphere*, *11*(5), 513.

Shuster, W. D., Bonta, J., Thurston, H., Warnemuende, E., & Smith, D. (2005). Impacts of impervious surface on watershed hydrology: A review. *Urban Water Journal*, *2*(4), 263–275.

Stagakis, S., Chrysoulakis, N., Spyridakis, N., Feigenwinter, C., & Vogt, R. (2019). Eddy covariance measurements and source partitioning of $CO_2$ emissions in an urban environment: Application for heraklion, greece. *Atmospheric Environment*, *201*, 278–292.

Steeneveld, G.-J., van der Horst, S., & Heusinkveld, B. (2020). Observing the surface radiation and energy balance, carbon dioxide and methane fluxes over the city centre of Amsterdam. In *EGU general assembly conference abstracts* (p. 1547).

Stewart, I. D., & Oke, T. R. (2012). Local climate zones for urban temperature studies. *Bulletin of the American Meteorological Society*, *93*(12), 1879–1900.

Sun, R., Wang, Y., & Chen, L. (2018). A distributed model for quantifying temporal-spatial patterns of anthropogenic heat based on energy consumption. *Journal of Cleaner Production*, *170*, 601–609.

Templeton, N. P., Vivoni, E. R., Wang, Z.-H., & Schreiner-McGraw, A. P. (2018). Quantifying water and energy fluxes over different urban land covers in phoenix, arizona. *Journal of Geophysical Research: Atmospheres*, *123*(4), 2111–2128.

Tewari, M., Chen, F., Kusaka, H., & Miao, S. (2007). Coupled WRF/Unified Noah/urban-canopy modeling system. In *Ncar WRF documentation, NCAR, Boulder* (Vol. 122, pp. 1–22). Citeseer.

Thatcher, M., & Hurley, P. (2012). Simulating australian urban climate in a mesoscale atmospheric numerical model. *Boundary-Layer Meteorology*, *142*(1), 149–175.

Twine, T. E., Kustas, W., Norman, J., Cook, D., Houser, P., Meyers, T., ... Wesely, M. (2000). Correcting eddy-covariance flux underestimates over a grassland. *Agricultural and forest meteorology*, *103*(3), 279–300.

United Nations. (2018). *World urbanization prospects, the 2018 revision.* UN Department of Economic and Social Affairs.

Velasco, E., Perrusquia, R., Jiménez, E., Hernández, F., Camacho, P., Rodríguez, S., ... Molina, L. (2014). Sources and sinks of carbon dioxide in a neighborhood of Mexico City. *Atmospheric Environment*, *97*, 226–238.

Velasco, E., Pressley, S., Grivicke, R., Allwine, E., Molina, L. T., & Lamb, B. (2011). Energy balance in urban Mexico City: observation and parameterization during the MILAGRO/MCMA-2006 field campaign. *Theoretical and applied climatology*, *103*, 501–517.

Vulova, S., Meier, F., Rocha, A. D., Quanz, J., Nouri, H., & Kleinschmit, B. (2021). Modeling urban evapotranspiration using remote sensing, flux footprints, and

artificial intelligence. *Science of The Total Environment*, *786*, 147293.
Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*(3), 426–482.
Walsh, C. J., Fletcher, T. D., & Ladson, A. R. (2005). Stream restoration in urban catchments through redesigning stormwater systems: looking to the catchment to save the stream. *Journal of the North American Benthological Society*, *24*(3), 690–705.
Wang, C., Wang, Z.-H., & Ryu, Y.-H. (2021). A single-layer urban canopy model with transmissive radiation exchange between trees and street canyons. *Building and Environment*, *191*, 107593.
Wang, Y. P., Kowalczyk, E., Leuning, R., Abramowitz, G., Raupach, M. R., Pak, B., ... Luhar, A. (2011). Diagnosing errors in a land surface model (cable) in the time and frequency domains. *Journal of Geophysical Research: Biogeosciences*, *116*(G1).
Wang, Z.-H., Bou-Zeid, E., & Smith, J. A. (2013). A coupled energy transport and hydrological model for urban canopies evaluated using a wireless sensor network. *Quarterly Journal of the Royal Meteorological Society*, *139*(675), 1643–1657.
Ward, H. C., Evans, J. G., & Grimmond, C. S. B. (2013). Multi-season eddy covariance observations of energy, water and carbon fluxes over a suburban area in Swindon, UK. *Atmospheric Chemistry and Physics*, *13*(9), 4645–4666.
Ward, H. C., Kotthaus, S., Järvi, L., & Grimmond, C. S. B. (2016). Surface urban energy and water balance scheme (SUEWS): development and evaluation at two UK sites. *Urban Climate*, *18*, 1–32.
Wenzel Jr, H. G., & Voorhees, M. L. (1981). *Evaluation of the urban design storm concept.* University of Illinois at Urbana-Champaign. Water Resources Center.
Willmott, C. J. (1982). Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society*, *63*(11), 1309–1313.
WMO. (2019). *Guidance on integrated urban hydrometeorological, climate and environmental services - volume i.* Geneva.
Wouters, H., Demuzere, M., De Ridder, K., & van Lipzig, N. P. (2015). The impact of impervious water-storage parametrization on urban climate modelling. *Urban Climate*, *11*, 24–50.
Yang, Z.-L., Dickinson, R., Henderson-Sellers, A., & Pitman, A. (1995). Preliminary study of spin-up processes in land surface models with the first stage data of project for intercomparison of land surface parameterization schemes phase 1 (a). *Journal of Geophysical Research: Atmospheres*, *100*(D8), 16553–16578.
Yao, L., Wei, W., & Chen, L. (2016). How does imperviousness impact the urban rainfall-runoff process under various storm cases? *Ecological indicators*, *60*, 893–905.
Yu, M., Wu, H., Yin, J., Liang, X., & Miao, S. (2022). Improved delineation of urban hydrological processes in coupled regional climate models. *Water Resources Research*, e2022WR032695.
Zeisl, P., Mair, M., Kastlunger, U., Bach, P. M., Rauch, W., Sitzenfrei, R., & Kleidorfer, M. (2018). Conceptual urban water balance model for water policy testing: an approach for large scale investigation. *Sustainability*, *10*(3), 716.
Zhou, Q., Leng, G., Su, J., & Ren, Y. (2019). Comparison of urbanization and climate change impacts on urban flood volumes: Importance of urban planning and drainage adaptation. *Science of the Total Environment*, *658*, 24–33.