Harmonized gap-filled datasets from 20 urban flux tower sites

Mathew Lipson1,2, Sue Grimmond2, Martin Best3, Winston Chow4, Andreas Christen5, Nektarios Chrysoulakis6, Andrew Coutts7, Ben Crawford8, Stevan Earl9, Jonathan Evans10, Krzysztof Fortuniak11, Bert G. Heusinkveld12, Je-Woo Hong13, Jinkyu Hong14, Leena Järvi15, Sungsoo Jo16, Yeon-Hee Kim17, Simone Kotthaus18, Keunmin Lee19, Valéry Masson20, Joseph P. McFadden21, Oliver Michels22, Wlodzimierz Pawlak23, Matthias Roth24, Hirofumi Sugawara25, Nigel Tapper26, Erik Velasco27, Helen Claire Ward28

1Australian Research Council (ARC) Centre of Excellence for Climate System Science, Climate Change Research Centre, Level 4, Mathews Building, UNSW Sydney, New South Wales, 2052, Australia.
2Department of Meteorology, University of Reading, Reading, RG6 6ET, United Kingdom
3Met Office, Fitzroy Road, Exeter, Devon, EX1 3PB, United Kingdom.
4School of Social Sciences, Singapore Management University, Singapore.
5Environmental Meteorology, Institute of Earth and Environmental Sciences, Faculty of Environment and Natural Resources, University of Freiburg, Freiburg, Germany.
7School of Earth, Atmospheric and Environment, Monash University, Melbourne, Australia.
8Geography and Environmental Sciences, University of Colorado, Denver, Colorado, USA.
9Global Institute of Sustainability and Innovation, Arizona State University, Tempe, Arizona, USA.
10UK Centre for Ecology & Hydrology, Crowmarsh Gifford, Wallingford, United Kingdom.
11Department of Meteorology and Climatology, University of Lodz, Lodz, Poland.
12Department of Meteorology and Air Quality, Wageningen University, The Netherlands.
13Korea Environment Institute, Sejong, Republic of Korea.
14Ecosystem-Atmosphere Process Lab, Department of Atmospheric Sciences, Yonsei University, Seoul, Korea (Republic of).
15Institute for Atmospheric and Earth System Research/Physics, Faculty of Science, University of Helsinki, Helsinki, Finland.
16Ecosystem-Atmosphere Process Lab, Department of Atmospheric Sciences, Yonsei University, Seoul, Korea (Republic of).
17National Institute of Meteorological Sciences, Korea Meteorological Administration, Jeju, Korea (Republic of).
18Institut Pierre Simon Laplace (IPSL), CNRS, École Polytechnique, Institut Polytechnique de Paris, 91128 Palaiseau Cedex, France.
19Ecosystem-Atmosphere Process Lab, Department of Atmospheric Sciences, Yonsei University, Seoul, Republic of Korea.
20Centre National de Recherches Météorologiques, University of Toulouse, Météo-France and CNRS, Toulouse, France.
21Department of Geography and Earth Research Institute, University of California, Santa Barbara, United States.
22Environmental Meteorology, Albert-Ludwigs-University, Freiburg, Germany.
23Department of Meteorology and Climatology, University of Lodz, Lodz, Poland.
24Department of Geography, National University of Singapore, Singapore.
26School of Earth, Atmospheric and Environment, Monash University, Melbourne, Australia.
27Independent Research Scientist, Singapore.
28Department of Atmospheric and Cryospheric Sciences, University of Innsbruck, Innsbruck, Austria.

Correspondence to: Mathew J. Lipson (mathew.lipson@unsw.edu.au)
Abstract. Twenty urban neighbourhood-scale eddy covariance flux tower datasets have been harmonized and quality controlled, producing a 50 site-year collection with broad diversity in climate and urban surface characteristics. Observations are gap-filled and prepended with 10 years of reanalysis-derived local data to enable use as spin up and forcing for land surface model evaluation. For both gap filling and spin-up, ERA5 reanalysis meteorological data are bias corrected using tower observations, accounting for diurnal, seasonal and local urban effects not modelled in ERA5. The bias correction methods developed perform well compared to methods used in other datasets (e.g. WFDE5 or FLUXNET2015). Site description metadata includes local land cover fractions (buildings, roads, trees, grass etc.), building height and morphology, aerodynamic roughness estimates, population density and satellite imagery. Together, this collection can help extend our understanding of urban environmental processes through observational synthesis studies or in the evaluation of land surface environmental models in a wide range of urban settings.

1 Background

Tower mounted instruments allow the measurement of land-atmosphere fluxes (e.g. energy, momentum, water, carbon) and local meteorological conditions. These observations are one of the fundamental ways of improving both our understanding and ability to predict biogeochemical and weather-related processes at local scales. Regional and global networks of flux tower sites have helped extend our knowledge of ecosystem and climate science (Novick et al., 2018; Beringer et al., 2016; Yamamoto et al., 2005; Valentini, 2003). Over the last 25 years networks such as FLUXNET have progressively increased access to flux data through open-source collections (Pastorello et al., 2020), extending the reach and impact of individual site observations through synthesis studies (Baldocchi, 2020) and multi-site environmental modelling and model evaluation projects (Best et al., 2015; Ukkola et al., 2022). However, with few urban sites included, urban areas have not benefited from the improved understanding or more extensive model evaluations that these collections can facilitate.

Urban areas are unique ecosystems, distinct from natural or rural landscapes. First, most people live in cities (UN, 2018) and infrastructure is concentrated within them. Therefore, climate-related health and economic impacts fall disproportionately within urban areas. Second, urban infrastructure (e.g. buildings and roads) along with transient human activities (e.g. energy consumption and irrigation) fundamentally alter surface energy, water and mass exchanges with the atmosphere, modifying local and larger-scale environmental conditions (Oke et al., 2017). Third, as built environments, urban areas are uniquely capable of actively mitigating and adapting to climate change.

Establishing and maintaining long term flux sites in cities is particularly challenging because of the rarity of appropriate sites with homogenous fetch, the difficulty in gaining approval to access existing towers (e.g. for telecommunications), the cost of constructing tall towers over an aerodynamically rough surface, and extremely limited long-term funding opportunities (Arnfield, 2003; Grimmond, 2006; Velasco and Roth, 2010; Feigenwinter et al., 2012; Grimmond and Ward, 2021). Thus, despite the diversity and importance of urban areas across the globe, urban flux tower data are relatively scarce, generally of...
short duration and rarely open source. Databases identifying urban observational programmes exist (e.g. the Urban Flux
Network (Grimmond and Christen, 2012)), however urban flux tower datasets have not previously been brought together into
a harmonised, gap-filled, open access collection.

We bring together quality-controlled data from 20 urban sites in an open collection that includes 50 observation-years (Lipson
et al., 2022). The sites are chosen to be diverse in both regional climates and urban characteristics. As evaluating land surface
models is one key application for these data, we create continuous forcing data sets (i.e. with incoming radiation fluxes and
other meteorological data) that are gap filled using site specific, bias corrected reanalysis data. Observations are also prepended
with 10 years of site-specific reanalysis-derived meteorological data to allow modelled soil moisture and other conditions to
equilibrante with local climate conditions during model spin up.

Along with the meteorological data, site characteristics and metadata are provided in a common format. The metadata includes
tower location, land cover fractions, building heights and morphology, aerodynamic roughness parameter estimates, population
density, estimated anthropogenic heat fluxes, site photos and satellite imagery. Together, this collection can help extend our
understanding of urban environmental processes through observational synthesis studies, or in the evaluation of land surface
environmental models in different urban settings.

2 Methods

2.1 Site selection

The initial motivation for collating these flux tower and site data is for use in the Urban-PLUMBER multi-site model evaluation
project, currently underway (Urban-PLUMBER: A multi-site model evaluation project for urban areas - Project Home, 2021).
Urban-PLUMBER draws on methods from the first international urban land surface model comparison (Grimmond et al.,
2010, 2011) and the Protocol for the Analysis of Land Surface Models Benchmarking Evaluation Project (PLUMBER (Best
et al., 2015)). The latter evaluated land surface models in non-urban (vegetated) areas, while Urban-PLUMBER evaluates land
surface models at 20 urban sites (Fig. 1, Table 1).

There is a two-fold use of these observational data in the model evaluation of Urban-PLUMBER:

1. To provide local-scale meteorological input forcing to drive land surface models

2. To evaluate the performance of models, primarily assessing the local-scale exchange of radiant and turbulent heat
fluxes between the surface and lower atmosphere

With these objectives in mind, the following criteria are used to select flux tower sites:

- Appropriately sited for neighbourhood-scale conditions - i.e. within the inertial sub-layer, typically 2-5 times above
the average building height and with relatively homogenous fetch (Grimmond, 2006; Barlow, 2014; Grimmond and
Ward, 2021)
• Requested observations available at 30- or 60-minute resolution (Table 2)
• Local site characteristics available for description and configuring models
• A preference for longer datasets (as this allows seasonal and inter-annual variability to be included)
• Collectively represent a diverse range of site characteristics and climates

Table 1: Site location and included observation (focus) period. Data providers may have longer observation periods available than are in this collection. Resolution is 30 minutes (or 60 minutes if denoted by *). All periods in universal time coordinated (UTC). US-Minneapolis data are split based on wind direction and fetch (Section 4.5).

<table>
<thead>
<tr>
<th>Sitename</th>
<th>City</th>
<th>Country</th>
<th>Observed period</th>
<th>Latitude</th>
<th>Longitude</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>AU-Preston</td>
<td>Melbourne</td>
<td>Australia</td>
<td>Aug 2003 – Nov 2004</td>
<td>-37.7306</td>
<td>145.0145</td>
<td>(Coutts et al., 2007a, b)</td>
</tr>
<tr>
<td>AU-SurreyHills</td>
<td>Melbourne</td>
<td>Australia</td>
<td>Feb 2004 – Jul 2004</td>
<td>-37.8265</td>
<td>145.099</td>
<td>(Coutts et al., 2007a, b)</td>
</tr>
<tr>
<td>CA-Sunset</td>
<td>Vancouver</td>
<td>Canada</td>
<td>Jan 2012 – Dec 2016</td>
<td>49.2261</td>
<td>-123.078</td>
<td>(Christen et al., 2011; Crawford and Christen, 2015)</td>
</tr>
<tr>
<td>FI-Kumpula</td>
<td>Helsinki</td>
<td>Finland</td>
<td>Dec 2010 – Dec 2013</td>
<td>60.2028</td>
<td>24.9611</td>
<td>(Karsisto et al., 2016)</td>
</tr>
<tr>
<td>FI-Torni</td>
<td>Helsinki</td>
<td>Finland</td>
<td>Dec 2010 – Dec 2013</td>
<td>60.1678</td>
<td>24.9387</td>
<td>(Järvi et al., 2018; Nordbo et al., 2013)</td>
</tr>
<tr>
<td>FR-Capitole</td>
<td>Toulouse</td>
<td>France</td>
<td>Feb 2004 – Mar 2005</td>
<td>43.6035</td>
<td>1.4454</td>
<td>(Masson et al., 2008; Goret et al., 2019)</td>
</tr>
<tr>
<td>JP-Yoyogi</td>
<td>Tokyo</td>
<td>Japan</td>
<td>Mar 2016 – Mar 2020*</td>
<td>35.6645</td>
<td>139.6845</td>
<td>(Hirano et al., 2015; Ishidoya et al., 2020)</td>
</tr>
<tr>
<td>KR-Jungnang</td>
<td>Seoul</td>
<td>South Korea</td>
<td>Jan 2017 – Apr 2019</td>
<td>37.5907</td>
<td>127.0794</td>
<td>(Jo et al., n.d.; Hong et al., 2020)</td>
</tr>
<tr>
<td>NL-Amsterdam</td>
<td>Amsterdam</td>
<td>Netherlands</td>
<td>Jan 2019 – Oct 2020</td>
<td>52.3665</td>
<td>4.8929</td>
<td>-</td>
</tr>
</tbody>
</table>
Potential sites identified from published site lists (Grimmond and Christen, 2012; Oke et al., 2017) and open calls for data (e.g. community newsletters (Lipson et al., 2020a), international conferences (Lipson et al., 2020b, c) and social media professional networks). We deemed 20 sites sufficient for the evaluation project (Table 1), together covering a 50 site-years. Included sites have built fractions (i.e. plan area fraction of all impervious surfaces including roofs, roads, other paving etc.) from 0.05 to 0.965, and are located in four major Köppen-Geiger (Beck et al., 2018) climate classes (Fig. 1). Eleven sites are in temperate climates, eight in cold (or continental) climates, and one in each of tropical and arid climates.

Sites are reasonably distributed across mean temperature and precipitation for global urban locations, but gaps remain, particularly in warm, wet and very cold climates (Fig. 2). Some urban flux observations in understudied regions were not included (e.g., Ouagadougou (Offerle et al., 2005), São Paulo (Ferreira et al., 2013), Guangzhou (Shi et al., 2019), Beijing (Dou et al., 2019)) because they do not meet the model evaluation project needs because of the relatively short observed periods for the available data. These regions and climates have large urban populations with significant environmental challenges and have few urban flux tower sites compared with northern hemisphere temperate or continental locations (Grimmond, 2006; Roth et al., 2017). Understudied regions and climates should be included in future collections when appropriate time series become available.

<table>
<thead>
<tr>
<th>Site</th>
<th>City</th>
<th>Country</th>
<th>Start Date – End Date*</th>
<th>Lat</th>
<th>Long</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>PL-Lipowa</td>
<td>Łódź</td>
<td>Poland</td>
<td>Jan 2008 – Dec 2012</td>
<td>51.7625</td>
<td>19.4453</td>
<td>(Fortuniak et al., 2013; Pawlak et al., 2011)</td>
</tr>
<tr>
<td>SG-TelokKurau06</td>
<td>Singapore</td>
<td>Singapore</td>
<td>Apr 2006 – Mar 2007</td>
<td>1.3143</td>
<td>103.9112</td>
<td>(Roth et al., 2017)</td>
</tr>
<tr>
<td>UK-KingsCollege</td>
<td>London</td>
<td>UK</td>
<td>Apr 2012 – Jan 2014</td>
<td>51.5118</td>
<td>-0.1167</td>
<td>(Bjorkegren et al., 2015; Kotthaus and Grimmond, 2014a, b)</td>
</tr>
<tr>
<td>UK-Swindon</td>
<td>Swindon</td>
<td>UK</td>
<td>May 2011 – Apr 2013</td>
<td>51.5846</td>
<td>-1.7981</td>
<td>(Ward et al., 2013)</td>
</tr>
<tr>
<td>US-Minneapolis</td>
<td>Minneapolis</td>
<td>USA</td>
<td>Jun 2006 – May 2009</td>
<td>44.9984</td>
<td>-93.1884</td>
<td>(Peters et al., 2011; Menzer and McFadden, 2017)</td>
</tr>
<tr>
<td>US-WestPhoenix</td>
<td>Phoenix</td>
<td>USA</td>
<td>Dec 2011 – Jan 2013</td>
<td>44.9984</td>
<td>-93.1884</td>
<td>(Chow, 2017; Chow et al., 2014)</td>
</tr>
</tbody>
</table>
Figure 2: Climatology of included sites compared with more than 70,000 global urban areas. Mean temperature and annual precipitation at the 20 tower sites (red, truncated site name, Table 1) from tower observations; global urban locations (grey) from ERA5 surface data (Hersbach et al., 2020, 2018) (2000 – 2010) from grid nearest to locations identified in the Global Rural-Urban Mapping Project (GRUMP) (Center for International Earth Science Information Network - CIESIN - Columbia University et al., 2017). Locations with rainfall above 3000 mm year-1 (1.3% of locations), and mean temperature below -3°C (0.2% of locations) are not shown.
2.2 Flux tower data

The observed data are provided in 30- or 60-minute periods (Table 1), processed from high-frequency samples by individual observing groups. In the harmonized collection time stamps are in coordinated universal time (UTC) indicating the end of the measurement period. Variables name and units use ALMA conventions: (Assistance for Land-surface Modelling Activities), a format used in previous land surface model comparisons.

Data are cleaned (Sect 2.3: Quality control), gap filled (Section 2.4) and prepended with data derived from ERA5 (Section 2.5) after site-specific corrections (Section 2.6). An example of final prepended and gap-filled data is shown in Figure 3 for one site (UK-KingsCollege). Plots for all other variables and sites are also available in the collection (Lipson et al., 2022).

Data are split into forcing and analysis variable sets (Table 2) to allow the forcing variables to be provided to modelling groups as input to run their models. The withheld analysis data are used by the coordinating group to assess the model outputs.

Some additional observed variables (Table 3) have, where practical, been included in the datasets after passing through the quality control steps. Missing forcing variables are obtained using bias-corrected reanalysis data (Section 2.4). No gap filling is applied to analysis data or additional variables.
Figure 3: Forcing timeseries with gap filling. Example shown for air temperature (Tair) at Kings College, London (UK-KingsCollege); (a) full forcing period, including 10 years of ERA5-derived data (red) prior to observations (black) used for model spin up, and (b) focus period used for analysis. Gaps are first filled from nearby tower measurements where available, and short gaps (≤2 hours) are linearly interpolated (blue). Remaining gaps are filled using the ERA5-derived timeseries which is seasonally and diurnally bias corrected using site observations. White lines show seven day mean values. Similar plots are available for other sites within the site data collection (Lipson et al., 2022).

Table 2: Forcing and analysis flux tower data variables. Short name description, units and positive direction use ALMA data conventions. Mean annual estimates of anthropogenic heat flux are included as site metadata. Note ground heat flux (Qg) is the heat flux into soil rather than total storage heat flux which is difficult to measure in urban areas (Grimmond and Oke, 1999).
## Table 3: Site climate classification, missing and additional variables. Climate classification from Köppen-Geiger global dataset (Beck et al., 2018). Table 2 gives variable definitions. Note: as MX-Escandon LWdown data are unavailable during the 2011-2012 focus period, but were available in 2006, this earlier period is used to determine bias correction for ERA5 LWdown data.

<table>
<thead>
<tr>
<th>SiteName</th>
<th>Class</th>
<th>Climate description</th>
<th>Missing variables</th>
<th>Additional variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>AU-Preston</td>
<td>Cfb</td>
<td>Temperate, no dry season, warm summer</td>
<td>Snowf</td>
<td>Qtau</td>
</tr>
<tr>
<td>AU-SurreyHills</td>
<td>Cfb</td>
<td>Temperate, no dry season, warm summer</td>
<td>Snowf</td>
<td>Qtau</td>
</tr>
<tr>
<td>CA-Sunset</td>
<td>CsB</td>
<td>Temperate, dry summer, warm summer</td>
<td>Snowf</td>
<td>Qtau, SoilTemp</td>
</tr>
<tr>
<td>FI-Kumpula</td>
<td>Dfb</td>
<td>Cold, no dry season, warm summer</td>
<td>Snowf</td>
<td></td>
</tr>
<tr>
<td>FI-Torni</td>
<td>Dfb</td>
<td>Cold, no dry season, warm summer</td>
<td>Snowf</td>
<td></td>
</tr>
<tr>
<td>FR-Capitole</td>
<td>Cfa</td>
<td>Temperate, no dry season, hot summer</td>
<td>Snowf</td>
<td>Qtau</td>
</tr>
<tr>
<td>GR-HECKOR</td>
<td>Csa</td>
<td>Temperate, dry summer, hot summer</td>
<td>Snowf</td>
<td>Qtau</td>
</tr>
<tr>
<td>JP-Yoyogi</td>
<td>Cfa</td>
<td>Temperate, no dry season, hot summer</td>
<td>Snowf</td>
<td>Qtau</td>
</tr>
<tr>
<td>KR-Jungnang</td>
<td>Dwa</td>
<td>Cold, dry winter, hot summer</td>
<td>Snowf</td>
<td></td>
</tr>
<tr>
<td>KR-Ochang</td>
<td>Dwa</td>
<td>Cold, dry winter, hot summer</td>
<td>Snowf</td>
<td></td>
</tr>
<tr>
<td>MX-Escandon</td>
<td>Cwb</td>
<td>Temperate, dry winter, warm summer</td>
<td>Snowf, LWdown*</td>
<td>Qtau</td>
</tr>
<tr>
<td>NL-Amsterdam</td>
<td>Cfb</td>
<td>Temperate, no dry season, warm summer</td>
<td>Snowf</td>
<td>Qtau</td>
</tr>
<tr>
<td>PL-Lipowa</td>
<td>Dfb</td>
<td>Cold, no dry season, warm summer</td>
<td>Snowf</td>
<td></td>
</tr>
<tr>
<td>PL-Narutowicza</td>
<td>Dfb</td>
<td>Cold, no dry season, warm summer</td>
<td>Snowf</td>
<td></td>
</tr>
<tr>
<td>SG-TelokKurau06</td>
<td>Af</td>
<td>Tropical, rainforest</td>
<td>Snowf</td>
<td></td>
</tr>
<tr>
<td>UK-KingsCollege</td>
<td>Cfb</td>
<td>Temperate, no dry season, warm summer</td>
<td>Snowf</td>
<td>Qtau</td>
</tr>
<tr>
<td>UK-Swindon</td>
<td>Cfb</td>
<td>Temperate, no dry season, warm summer</td>
<td>Snowf</td>
<td>Qtau</td>
</tr>
<tr>
<td>US-Baltimore</td>
<td>Cfa</td>
<td>Temperate, no dry season, hot summer</td>
<td>Snowf</td>
<td>Qtau, SoilTemp</td>
</tr>
<tr>
<td>US-Minneapolis</td>
<td>Dfa</td>
<td>Cold, no dry season, hot summer</td>
<td>Snowf</td>
<td>Qtau, SoilTemp, Qg</td>
</tr>
<tr>
<td>US-WestPhoenix</td>
<td>BWh</td>
<td>Arid, desert, hot</td>
<td>Snowf</td>
<td>Qtau, SoilTemp</td>
</tr>
</tbody>
</table>

### 2.3 Quality control and assurance

For each site the 30- or 60-minute variables are calculated by data providers from high-frequency samples after applying their own quality control measures (e.g. Aubinet et al., 2012; Feigenwinter et al., 2012; Kotthaus and Grimmond, 2012; Vitale et al., 2020). The harmonised collection consists of the data retained after undergoing five additional quality control steps, in this order:
1. **Out-of-range**: removal of unphysical values (e.g., negative shortwave radiation) using the ALMA expected range protocol (Bowling and Polcher, 2001).

2. **Night**: nocturnal shortwave radiation set to zero, based on civil twilight (when the sun is $6^\circ$ below the horizon (Forsythe et al., 1995)).

3. **Constant**: four or more timesteps with identical values (excluding zero values for shortwave radiation, rainfall and snowfall) are removed as suspicious.

4. **Outlier**: values outside $\pm 4$ standard deviations for each hour in a rolling 30-day window (to account for diurnal and seasonal variations) removed. Repeat with a larger tolerance ($\pm 5$ standard deviations) until no outliers remain (Schmid et al., 2000). The outlier test is not applied to precipitation.

5. **Visual**: remaining suspect readings are removed manually via visual inspection.

These steps are undertaken in the processing script `qc_observations.py` (see Sect 5: Code availability), including periods identified through visual inspection (21 instances across all data). Data removed through quality control are indicated in plots of each variable at each site included in the data collection (Lipson et al., 2022). Unaltered (raw) data are also included in the collection.

Communication, or human errors, also have the potential to degrade or invalidate data (Menard et al., 2021). As part of quality assurance, project coordinators prepared an observational data protocol (Lipson et al., 2021) to explicitly set out requirements for data providers prior to submission of their data. The protocol documented instrument siting requirements, variables and data formats, dataset length and resolution, necessary site characteristic information and metadata, as well as the expectations for data handling, use and authorship. On receiving data, coordinators undertook further checks and identified errors that were not be picked up by automated quality control. Identified errors included mislabelled variables and metadata, inconsistent timestamps and unit discrepancies. Many of the errors were identified by comparing provided data with secondary sources such as ERA5, nearby meteorological stations or previous publications. Errors were corrected collaboratively with data providers, some leading to corrections in primary data sources.

### 2.4 Gap-filling

Three gap filling methods are used to create a continuous dataset for forcing variables, in this order:

- contemporaneous and nearby flux tower or weather observing sites (where available from data providers)
- small gaps ($\leq 2$ hours) are linearly interpolated from the adjoining observations
- larger gaps and a 10-year spin-up period are filled with bias corrected ERA5 data (Section 2.6).

As only one site provided observed snowfall rate (JP-Yoyogi), ERA5 snowfall rates are used for all periods at other sites. At those sites the additional water equivalent from ERA5 snowfall is removed from subsequent observed rainfall until mass balance of observed total precipitation is achieved. This corrects melting snow being recorded as rainfall.
2.5 ERA5 reanalysis data

The ERA5 reanalysis product (Hersbach et al., 2020) assimilates global satellite, atmospheric and ground-based observations to constrain numerical weather prediction simulations, producing global output at 0.25° spatial and hourly temporal resolutions from 1979 to the present. It is therefore useful as a globally consistent and accessible source of meteorological data across space and time. ERA5, and its lower resolution predecessor ERA-Interim (Dee et al., 2011), have been used extensively to provide meteorological forcing data to drive land surface models and gap fill flux tower observations (Vuichard and Papale, 2015; Kokkonen et al., 2018; Pastorello et al., 2020; Ukkola et al., 2017, 2022).

The ERA5 hourly single level (Hersbach et al., 2018) dataset (retrieved from NCI Australia (Druken, 2020)) is used for gap-filling missing observations within the focus periods (Table 1) and for the 10-year model spin-up period. However, combining ERA5 data directly with urban flux tower observations has several deficiencies.

Grid-scale ERA5 data are not directly compatible with point-scale urban flux tower observations. This incompatibility is three-fold:

1. **Horizontally**: The ERA5 grid cell area (of order 30 x 30 km²) does not match the flux footprint from tower observations (of order 1 km²). The ERA5 surface characteristics, including elevation, are based on an average description for the grid which may differ from surface characteristics around the observing tower, particularly in coastal or mountainous regions (Martens et al., 2020) in which many cities are located.

2. **Vertically**: ERA5 provide near-surface variables (2 or 10 m above ground level), aligning with World Meteorological Organization (WMO) guidelines for standard regional observations taken over short grass (World Meteorological Organization, 2008). As the urban roughness elements (e.g. buildings) are much taller than grass, instruments are mounted on towers at heights greater than 2 – 5 times average building height in order to be located within the inertial sub layer or constant flux layer (Velasco and Roth, 2010; Barlow, 2014; Grimmond and Ward, 2021).

3. **Land surface**: As the current operational ERA5 modelling systems do not include an urban land surface scheme (Boussetta et al., 2013; McNorton et al., 2021), other land types (grass, crops, shrubs, trees etc.) are to characterise the grid cell (Table 4). Urban land surfaces are well known to alter local meteorological conditions (Oke et al., 2017), therefore ERA5 output will likely differ from locally observed conditions.

Outside of cities there are known diurnal and seasonal biases between the ERA5 near-surface variables and observations (Haiden et al., 2018; Betts et al., 2019; Nogueira, 2020; Martens et al., 2020; Jiang et al., 2021). These biases are an outcome of simplifying assumptions made in model parameterisations and inadequacies of modelling frameworks in general (Cucchi et al., 2020). Various approaches to reduce ERA5 biases in non-urban areas have been proposed. For example, the Water and Global Change (WATCH) Forcing Data (WFD) project use gridded observations to bias-correct ERA-Interim data (Weedon et al., 2011), and more recently ERA5 data, creating the global WFDE5 dataset for impact studies (Cucchi et al., 2020). WFDE5
relies on the Climate Research Unit (CRU) monthly timeseries of gridded observations with resolution courser than ERA5 (New et al., 1999), requiring ERA5 to be regridded to a lower resolution. This may reduce the representativeness of the ERA5 data, particularly in heterogeneous or complex terrain.

Alternatively, local observations can be used to bias-correct ERA data, e.g. the linear regression corrections using tower observations applied to FLUXNET datasets (Vuichard and Papale, 2015; Pastorello et al., 2020). However, linear methods neither conserve the variability of observations (Vuichard and Papale, 2015) (Section 3), nor can they correct diurnal timing differences within ERA5 data (e.g. out of phase from urban temporal profiles, which are typically delayed compared with non-urban surfaces used in ERA5, Figure 4, Table 4).

To account for these deficiencies, we develop a novel set of methods to bias correct ERA5 data to better represent observed urban conditions (Section 2.6).

Table 4: ERA5 surface cover information for grid cells in which the urban flux towers are located (Table 1). The land surface model used for ERA5 simulations does not account for urbanised land surfaces, nor vegetation types typical at urban sites. ERA5 surface roughness values are indicative as they vary slightly through time. Effective roughness is a correction accounting for observed urban mean wind speeds.

<table>
<thead>
<tr>
<th>Site</th>
<th>ERAS low vegetation</th>
<th>ERAS high vegetation</th>
<th>low vegetation fraction</th>
<th>high vegetation fraction</th>
<th>lake (or sea) fraction</th>
<th>bare soil fraction</th>
<th>effective roughness [m]</th>
<th>surface roughness [m]</th>
</tr>
</thead>
<tbody>
<tr>
<td>AU-Preston</td>
<td>tall grass</td>
<td>interrupted forest</td>
<td>0.484</td>
<td>0.407</td>
<td>0.088</td>
<td>0.021</td>
<td>0.514</td>
<td>0.289</td>
</tr>
<tr>
<td>AU-SurreyHills</td>
<td>tall grass</td>
<td>interrupted forest</td>
<td>0.484</td>
<td>0.407</td>
<td>0.088</td>
<td>0.021</td>
<td>0.514</td>
<td>0.368</td>
</tr>
<tr>
<td>CA-Sunset</td>
<td>crops, mixed farming</td>
<td>evergreen needleleaf trees</td>
<td>0.205</td>
<td>0.723</td>
<td>0.071</td>
<td>0.000</td>
<td>1.077</td>
<td>1.508</td>
</tr>
<tr>
<td>FI-Kumpula</td>
<td>crops, mixed farming</td>
<td>evergreen needleleaf trees</td>
<td>0.296</td>
<td>0.352</td>
<td>0.137</td>
<td>0.215</td>
<td>0.708</td>
<td>0.703</td>
</tr>
<tr>
<td>FI-Torni</td>
<td>crops, mixed farming</td>
<td>evergreen needleleaf trees</td>
<td>0.296</td>
<td>0.352</td>
<td>0.137</td>
<td>0.215</td>
<td>0.708</td>
<td>0.424</td>
</tr>
<tr>
<td>FR-Capitole</td>
<td>crops, mixed farming</td>
<td>interrupted forest</td>
<td>0.920</td>
<td>0.050</td>
<td>0.004</td>
<td>0.025</td>
<td>0.291</td>
<td>0.519</td>
</tr>
<tr>
<td>GR-HECKOR</td>
<td>crops, mixed farming</td>
<td>interrupted forest</td>
<td>0.172</td>
<td>0.463</td>
<td>0.158</td>
<td>0.207</td>
<td>0.505</td>
<td>1.187</td>
</tr>
<tr>
<td>JP-Yoyogi</td>
<td>semidesert</td>
<td>no vegetation recorded</td>
<td>0.943</td>
<td>0.000</td>
<td>0.010</td>
<td>0.047</td>
<td>0.015</td>
<td>0.649</td>
</tr>
<tr>
<td>KR-Jungnang</td>
<td>crops, mixed farming</td>
<td>evergreen needleleaf trees</td>
<td>0.781</td>
<td>0.168</td>
<td>0.051</td>
<td>0.000</td>
<td>0.516</td>
<td>0.074</td>
</tr>
<tr>
<td>KR-Ochang</td>
<td>irrigated crops</td>
<td>interrupted forest</td>
<td>0.281</td>
<td>0.716</td>
<td>0.003</td>
<td>0.000</td>
<td>0.844</td>
<td>0.181</td>
</tr>
<tr>
<td>MX-Escandon</td>
<td>evergreen shrubs</td>
<td>mixed forest/woodland</td>
<td>0.743</td>
<td>0.216</td>
<td>0.006</td>
<td>0.035</td>
<td>0.404</td>
<td>0.229</td>
</tr>
<tr>
<td>NL-Amsterdam</td>
<td>crops, mixed farming</td>
<td>interrupted forest</td>
<td>0.867</td>
<td>0.061</td>
<td>0.056</td>
<td>0.015</td>
<td>0.248</td>
<td>0.254</td>
</tr>
<tr>
<td>PL-Lipowa</td>
<td>crops, mixed farming</td>
<td>interrupted forest</td>
<td>0.855</td>
<td>0.144</td>
<td>0.001</td>
<td>0.000</td>
<td>0.250</td>
<td>0.306</td>
</tr>
<tr>
<td>PL-Narutowicza</td>
<td>crops, mixed farming</td>
<td>interrupted forest</td>
<td>0.855</td>
<td>0.144</td>
<td>0.001</td>
<td>0.000</td>
<td>0.250</td>
<td>0.550</td>
</tr>
<tr>
<td>SG-TelokKuruau06</td>
<td>irrigated crops</td>
<td>interrupted forest</td>
<td>0.905</td>
<td>0.021</td>
<td>0.074</td>
<td>0.000</td>
<td>0.335</td>
<td>0.309</td>
</tr>
<tr>
<td>UK-KingsCollege</td>
<td>crops, mixed farming</td>
<td>interrupted forest</td>
<td>0.609</td>
<td>0.372</td>
<td>0.020</td>
<td>0.000</td>
<td>0.504</td>
<td>0.315</td>
</tr>
<tr>
<td>UK-Swindon</td>
<td>crops, mixed farming</td>
<td>interrupted forest</td>
<td>0.727</td>
<td>0.251</td>
<td>0.001</td>
<td>0.021</td>
<td>0.397</td>
<td>0.146</td>
</tr>
<tr>
<td>US-Baltimore</td>
<td>crops, mixed farming</td>
<td>deciduous broadleaf trees</td>
<td>0.044</td>
<td>0.908</td>
<td>0.048</td>
<td>0.000</td>
<td>1.675</td>
<td>1.076</td>
</tr>
<tr>
<td>US-Minneapolis1</td>
<td>crops, mixed farming</td>
<td>interrupted forest</td>
<td>0.228</td>
<td>0.706</td>
<td>0.059</td>
<td>0.006</td>
<td>0.814</td>
<td>0.242</td>
</tr>
<tr>
<td>US-Minneapolis2</td>
<td>crops, mixed farming</td>
<td>interrupted forest</td>
<td>0.228</td>
<td>0.706</td>
<td>0.059</td>
<td>0.006</td>
<td>0.814</td>
<td>0.406</td>
</tr>
<tr>
<td>US-WestPhoenix</td>
<td>semidesert</td>
<td>evergreen needleleaf trees</td>
<td>0.949</td>
<td>0.051</td>
<td>0.000</td>
<td>0.000</td>
<td>0.084</td>
<td>0.404</td>
</tr>
</tbody>
</table>

2.6 Bias-correction methods

Bias-correction approaches used in the collection depend on the forcing variable (Table 2) and are described below.
2.6.1 Hourly and daily corrections

For incoming longwave radiation, air temperature, specific humidity and air pressure, the mean bias between ERA5 and local flux tower observations are calculated for each hour \( (h) \) and each day of a year \( (D) \) in a 60-day rolling window of a representative year (Fig. 4a). The calculated bias \( \eta_{bias}(D, h) \) is subtracted from the complete ERA5 timeseries \( \eta_{ERA5}(t) \) to create a new corrected timeseries:

\[
\eta(t) = \eta_{ERA5}(t) - \eta_{bias}(D, h) .
\]  

1. If observations have a 30-min resolution, average to 60-min to match ERA5 periods
2. Remove ERA5 data for time periods where site observations are missing
3. Calculate mean for each hour with both data sets to create a ‘representative’ year of 366 days
4. Extend to a three-year period by duplication to provide smoother transitions at year end
5. Calculate hourly means in a 60-day rolling window across the repeating timeseries, excluding data in windows with less than 30 observations. Repeat mean calculation for greater smoothing.
6. Calculate a timeseries of the bias between observed and ERA5 rolling means
7. Fill gaps in the bias timeseries by linear interpolation through each hour separately
8. Remove first and last year in the bias timeseries, using only the central year to bias correct each hour in the original ERA5 timeseries

![Figure 4: Urban-PLUMBER reanalysis bias correction methods. Demonstrated using air temperature (Tair) for the grid containing the King’s College London site (UK-KingsCollege). (a) hourly (colour) bias calculated for each day of a ‘representative’ year and applied to entire ERA5 timeseries; (b) diurnal hourly mean Urban-PLUMBER correction (UP, red), observations (black), original ERA5 data (blue),](https://doi.org/10.5194/essd-2022-65)
WFDE5 bias corrected data (green) and linear bias correction method used in FLUXNET (LN, yellow). Our new UP method has smaller mean absolute errors (MAE) overall, and can correct both pattern and phase errors of ERA5 (Section 3).

2.6.2 Logarithmic wind profile correction

Wind speed differences between ERA5 and site observations can result from errors in modelled synoptic-scale speeds, differences in representative heights, and differences in surface aerodynamic properties like roughness and displacement height. To correct bias and maintain standard deviations of the wind components (U,V) of observations at sensor height (\(z_{site}\)), the following correction to ERA5 data is undertaken assuming both a logarithmic wind profile and neutral conditions (Goret et al., 2019):

\[
u_{corr} = \frac{\ln(z_{site} - d_{site})}{\ln(z_{grid} - d_{grid})} \cdot u_{ERA5}
\]

where \(u_{corr}\) is the corrected wind speed at \(z_{site}\). The ERA5 wind (\(u_{ERA5}\)) at 10 m (\(z_{grid}\)) is used with the site surface roughness (\(z_{0,site}\)) and displacement height (\(d_{site}\)), grid roughness (\(z_{0,grid}\)) (Table 4) while assuming grid displacement height (\(d_{grid}\)) is zero for simplicity. If the resulting mean value of \(u_{corr}\) differs from observed mean value by more than 0.01 m s\(^{-1}\), then \(z_{0,grid}\) is iteratively adapted until this threshold accuracy in mean wind speed is achieved. Derived \(z_{0,grid}\) values are given in Table 4 (last column). Note this approach ignores seasonal effects from vegetation phenology and directional effects but ensures mean wind speeds are appropriate at the urban \(z_{site}\) while conserving variability within the ERA5 derived wind data.

2.6.3 Long term precipitation correction

Total precipitation is an important variable in urban land surface models because of the effect on soil moisture which evolves over multi-year periods (Best and Grimmond, 2014). Most of the observational datasets included are not long enough to capture interannual variations. We therefore use longer term total precipitation (\(P\)) from nearby stations from the Global Historical Climatology Network - Daily (GHCND) (Menne et al., 2012) over a 10 year period to correct ERA5 rain and snow fluxes (\(\phi\)) at each timestep:

\[
\phi_{corr} = \sum_{t=1}^{10\text{ yrs}} \frac{p_{GHCND}}{p_{ERA5}} \phi_{ERA5}(t).
\]

Gaps in the nearest GHCND station data are progressively filled by the next nearest station until no gaps are present. If gaps could not be filled with GHCND stations within 2° of latitude and longitude from the flux tower, and no alternative records are found (e.g. from national meteorological and hydrological services), then ERA5 rates are used unadjusted (viz., KR-Ochang and KR-Jungnang). This assumes precipitation occurs on the same dates and at the same times in ERA5 and observed datasets, which may become less valid under increasingly convective conditions.
2.6.4 Linear bias correction

The FLUXNET2015 collection of 212 flux tower sites (Pastorello et al., 2020) bias correction method uses linear regression between site observations and reanalysis data to derive one slope ($s$) and intercept ($b$) per site, hence ‘unbiasing’ all ERA timesteps ($i$):

$$LN_i = s \cdot ERA5_i + b. \quad (4)$$

Following FLUXNET2015, global radiation and wind fields are assigned an intercept of zero, and precipitation is not linearly modified (Vuichard and Papale, 2015). FLUXNET2015 use the coarser resolution ERA-Interim (spatial: 0.5° cf. 0.25°; temporal 3-h cf. 1-h) than ERA5. In this evaluation we use ERA5, which is found to be a consistent improvement over ERA-Interim (Albergel et al., 2018). However, after assessment we chose not to use a linear method (LN) for correcting variables in this collection. Its description is retained here for comparison purposes (Section 3).

3 Gap-filling evaluation

Site observations are quality controlled by individual data providers and collectively for this project (Section 2.3). The observed data required for forcing land surface models are then gap filled using a novel method of bias correcting reanalysis data. In this section, four methods which draw on ERA5 data are evaluated:

1. **ERA5**: nearest land based 0.25° resolution ERA5 (Hersbach et al., 2018, p.5) grid without bias correction
2. **W5**: nearest WFDE5 (Cucchi et al., 2020, p.5) grid (which uses bias correction from 0.5° CRU monthly gridded observations)
3. **UP**: the Urban-PLUMBER methods described here (using site observations for bias correction)
4. **LN**: linear methods based on FLUXNET2015 (Vuichard and Papale, 2015; Pastorello et al., 2020) (using site observations for bias correction)

To evaluate the methods available for gap filling, quality-controlled tower site observations ($O_i$) are used to assess the calculated value ($\eta$) at timestep $i$ using three metrics:

a) Mean bias error (MBE): $\frac{\sum_{i=1}^{n}(\eta_i - O_i)}{n}$

b) Mean absolute error (MAE): $\frac{\sum_{i=1}^{n}|\eta_i - O_i|}{n}$

c) Normalised standard deviation (nSD): $\frac{\sum_{i=1}^{n}(\eta_i - \bar{\eta})^2}{n-1} / \left( \frac{\sum_{i=1}^{n}(O_i - \bar{\eta})^2}{n-1} \right)$

where $\bar{\eta}$ and $\bar{O}$ are time-averaged over $n$ data points. The time stamps for each variable are made consistent between the data sources: ERA5 are hourly time ending data (Hersbach et al., 2018, p.5); 60-min site observations are hour ending; 30-min site
observations are converted to 60-min time ending by averaging; whereas as the WFDE5 SWdown, LWdown and Rainf are natively 60-min time beginning (Cucchi et al., 2020, p.5) they are shifted forward to match the time ending timestamps; and WFDE5 Tair, Qair, Psurf and Wind are instantaneous samples on the hour so their time stamp remains unchanged.

To summarise each metric for the 20 sites, we use boxplots (Fig. 5) for the seven forcing variables. The evaluation inherently considers the net differences associated with both the spatial (vertical and surface cover) differences and errors (model and observation) from the two datasets. Uncorrected ERA5 (blue, Fig. 5) biases are generally negative for Tair, LWdown and Wind, and generally positive for Qair and SWdown. These biases can be partly explained by the ERA5 framework not including an urban surface model. For example, the well-documented warmer air temperature in cities (urban heat island) are not modelled in ERA5 because natural land surfaces are assumed in simulations (Table 4), although ERA5 can include an urban signal if the data assimilated are from within urban areas (Tang et al., 2021). Qair shows a general positive bias as evapotranspiration will be overestimated in ERA5 without an urban land surface representation. Likewise, ERA5 SWdown are overestimated and LWdown underestimated possibly because urban air pollution effects are not included (Oke, 1988).

Other discrepancies between ERA5 data and site data arise from elevation differences in height above sea level (asl) of the ERA5 grid and site. For example, the MX-Escandon tower in Mexico City measurement height is 2277 m asl, whereas the ERA5 grid cell is assigned a surface elevation of 2540 m asl because the cell includes nearby mountains. This 263 m difference causes a negative bias to Psurf of 2594 Pa and contributes to a Tair difference of -2.15 K. Additionally, orographic uplift increases the grid cell rainfall, leading to positive ERA5 bias of 2.25 x 10^5 kg m^2 s^-1 (+710 mm year^-1) compared with the MX-Escandon site observations. The ERA5 rainfall bias is even more pronounced at CA-Sunset in Vancouver, Canada, with a +1178 mm year^-1 bias. These results are consistent with other studies highlighting discrepancies between reanalysis and local data in mountainous regions (Kokkonen et al., 2018).
Figure 5: Evaluation of bias correction methods. Four methods (colour) to create gap filled observed time series data: ERA5 (blue), WFDE5 (W5, green), linear debiasing (LN, orange), UP (red, this study) using (row 1) mean bias error, (row 2) mean absolute error, (row 3) normalised standard deviation, with the 20 individual sites (dots), and ideal agreement with observations (red line) and boxplot showing distribution. The UP corrections (selected for use in this study) have lower overall errors (cf. other methods) except SWdown, where no corrections to ERA5 are applied.

4 Data records

4.1 Data format

Timeseries data and site descriptive metadata are recorded in both plain text and netCDF4 (Rew et al., 1989) formats. Each site folder contains the following timeseries:

- [sitename]_raw_observations_[version]: site observed before project-wide quality control and gap filling (Table 1 gives period)
• [sitename]_clean_observations_[version]: after project-wide quality control and gap filling (Table 1 gives period)
• [sitename]_metforcing_[version]: continuous observations with reanalysis-derived data after quality control and gap filling (Table 2; forcing data for 10-year spin up, then Table 1 periods)
• [sitename]_era5_corrected_[version]: continuous timeseries (1990-2020) of bias corrected ERA5 reanalysis meteorological data (as used for gap filling and prepending metforcing observations)

Each site folder also contains the following site metadata:

• [sitename]_sitedata_[version].csv: comma separated text file for site characteristics metadata e.g. latitude, longitude, surface cover fraction, morphology etc. (Table 5, 6). This site characteristic data is also included within the metforcing netcdf (for convenience)
• index.html: A summary page of site information in html format, including site characteristics, site images, timeseries, gap filling, quality control and diurnal plots

4.2 Timeseries metadata

The timeseries files include the following metadata:

• title: short description of the file
• summary: longer description of the file
• sitename: site code (e.g. AU-Preston)
• long_sitename: site long name, including city and country information
• version: version of current file
• time_coverage_start: start of timeseries in UTC (includes spin-up)
• time_coverage_end: end of timeseries in UTC
• time_analysis_start: start of observed (focus) period in UTC
• time_shown_in: time standard (always UTC)
• local_utc_offset_hours: offset in hours of local time from UTC
• timestep_interval_seconds: period of block averaging in seconds (timestep)
• timestep_number_spinup: number of timesteps prior to observed focus period
• timestep_number_analysis: number of timesteps in observed focus period
• project_contact: contact details for the Urban-PLUMBER project coordinators
• observations_contact: contact details of the observational site data providers
• observations_reference: published references associated with the observations
• date_created: date and time of creation of this file
The netCDF files also include the following metadata for each variable

- **long_name**: plain language description of variable
- **standard_name**: equivalent variable name under the CF (climate and forecast) conventions
- **units**: SI (international system) units
- **ancillary_variables**: name of the associated quality control flag variable

NetCDF files also include site characteristics parameter values and descriptions (Table 5). Note, the local_utc_offset does not account for day light savings. As the onset/finishing of day light savings causes distinct behaviour changes it is critical users of the data consider this in their applications.

### 4.3 Site characteristics metadata

Site characteristics (Table 5, 6) are essential for any use of these data, and fundamental to application of land surface models. These metadata are provided in two machine readable forms (plain text in csv files, and netCDF4). The metadata are primarily drawn from published sources, or as advised by the data providers. If local parameters are not known values are estimated from high resolution global datasets or derived from empirical relations. The sources for each parameter are included within the site characteristic metadata.

There are numerous methods to estimate the probable extent and weighting of turbulent fluxes footprints relative to the eddy covariance sensors located on a flux tower (Velasco and Roth, 2010). The eddy covariance flux footprint provides a basis to identify which area (and weighting) should be used to estimate the land surface fractions impacting the measurements. Some studies in this collection determine the footprint and resulting land cover fractions dynamically (e.g. for each 30-min period), whereas others assumed a constant radius (e.g. based on a climatological analysis or rule of thumb) (Table 6). Standardising the method to determine land cover fractions across sites is beyond the scope of this work, so users of metadata should be mindful of these differences.

Different methods for estimating surface roughness length and zero-plane displacement height can give significantly different values (Kent et al., 2017). Given different sites have derived values using different methods we also provide values using two consistent morphometric methods (Macdonald (Macdonald et al., 1998) and Kanda (Kanda et al., 2013); Table 5, parameters 26 – 29) derived from surface fraction and building height parameters within the measurement footprint (Table 6). The Kanda modification to the Macdonald method accounts for the variability in roughness element height, resulting in larger displacement heights which are closer to estimates made with anemometric methods (Kent et al., 2017). The Macdonald method assumes that all the buildings have the same average roughness element height. However, care must be taken when using
Kanda values as some urban land surface models expect displacement height to be always lower than average building height (Hertwig et al., 2020).

Where not known, building height standard deviation (\(\sigma_H\)) is estimated from an empirical relation to building mean height (\(H_{\text{ave}}\)) (Kanda et al., 2013):

\[
\sigma_H = 1.05 H_{\text{ave}} - 3.7.
\]  
(5)

Similarly, unknown local wall to plan area ratios (\(\lambda_w\)) are derived from roof area fraction (\(\lambda_p\)) and canyon height to width ratio (\(H/W\)) assuming an infinite canyon geometry (Masson et al., 2020):

\[
\lambda_w = 2 (1 - \lambda_p) H/W.
\]  
(6)

Frontal area index (\(\lambda_f\)) is sometimes reported in site literature without \(H/W\) or \(\lambda_w\), in which case these are estimated (again assuming an infinite canyon geometry) with (Porson et al., 2010):

\[
\lambda_f = \frac{2}{\pi} (1 - \lambda_p) H/W.
\]  
(7)

Where not known or provided, mean annual anthropogenic heat flux (Varquez et al., 2021) or soil characteristics (Hengl, 2018a, b, c) are estimated from global datasets at 1 km or lower resolutions.

Table 5: Site characteristic metadata description and units. Select values are provided in Table 6.

<table>
<thead>
<tr>
<th>ID</th>
<th>Parameter</th>
<th>Units</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>latitude</td>
<td>degrees_north</td>
<td>Latitude of tower</td>
</tr>
<tr>
<td>2</td>
<td>longitude</td>
<td>degrees_east</td>
<td>Longitude of tower</td>
</tr>
<tr>
<td>3</td>
<td>ground_height</td>
<td>m</td>
<td>Height above sea level of base of tower</td>
</tr>
<tr>
<td>4</td>
<td>measurement_height_above_ground</td>
<td>m</td>
<td>Height above ground level (agl) of eddy covariance equipment on tower</td>
</tr>
<tr>
<td>5</td>
<td>impervious_area_fraction</td>
<td>1</td>
<td>Plan area fraction of all impervious (hard) surfaces, including roofs, roads, paths and paved areas</td>
</tr>
<tr>
<td>6</td>
<td>tree_area_fraction</td>
<td>1</td>
<td>Plan area fraction of tree canopy (&gt; 2 m)</td>
</tr>
<tr>
<td>7</td>
<td>grass_area_fraction</td>
<td>1</td>
<td>Plan area fraction of grass or other vegetation (&lt; 2 m)</td>
</tr>
<tr>
<td>8</td>
<td>bare_soil_area_fraction</td>
<td>1</td>
<td>Plan area fraction of bare soil</td>
</tr>
<tr>
<td>9</td>
<td>water_area_fraction</td>
<td>1</td>
<td>Plan area fraction of water</td>
</tr>
<tr>
<td>10</td>
<td>roof_area_fraction</td>
<td>1</td>
<td>Plan area fraction of roofs ((\lambda_p))</td>
</tr>
<tr>
<td>11</td>
<td>road_area_fraction</td>
<td>1</td>
<td>Plan area fraction of roads</td>
</tr>
<tr>
<td>12</td>
<td>other_paved_area_fraction</td>
<td>1</td>
<td>Plan area fraction of hard surfaces on ground excluding roads (e.g. paths, plazas, carparks etc)</td>
</tr>
<tr>
<td>13</td>
<td>building_mean_height</td>
<td>m</td>
<td>Mean height above ground of buildings ((H_{\text{ave}}))</td>
</tr>
<tr>
<td>14</td>
<td>tree_mean_height</td>
<td>m</td>
<td>Mean height above ground of trees</td>
</tr>
<tr>
<td>15</td>
<td>roughness_length_momentum</td>
<td>m</td>
<td>Aerodynamic roughness length for momentum as reported in literature or provided by data providers</td>
</tr>
<tr>
<td>16</td>
<td>displacement_height</td>
<td>m</td>
<td>Zero-plane displacement height as reported in literature or advised by data providers</td>
</tr>
<tr>
<td>17</td>
<td>canyon_height_width_ratio</td>
<td>1</td>
<td>Mean building height to mean street canyon width (distance between buildings) ratio ((H/W))</td>
</tr>
</tbody>
</table>
Table 6: Select site characteristic values (see Table 5 for definitions). Other site characteristic values and sources are provided within the collection (Lipson et al., 2022). Areas analysed for land cover fractions and roughness parameters are based on either a static radius around the flux tower (value given) or a dynamic footprint model (fpm), depending on the available site literature.
4.4 Data flags

Each variable for each timestep has a quality control (qc) flag. For example, LWdown_qc lists qc flags for LWdown at each timestep. Flag numbers are consistent across all variables:

0. observed by measurement at site and passes project quality control tests
1. filled by observation: interpolated from site observations over short (2 h) periods OR filled by observations from nearby (< 10 km) stations over longer periods
2. filled by ERA5: derived from ERA5 with site specific bias correction
3. missing or removed through quality control (occurs only in timeseries without gap filling)

4.5 Wind sector exclusions

Turbulent flux data are excluded from certain wind directions (Table 7) because of:

- interference on flow from tower structure (as identified by data providers)
- markedly different land cover characteristics from sectors of interest (with guidance from data providers)

The US-Minneapolis site has different surface cover by wind direction but is retained in the collection because of its both a long observation period and its distinct land cover characteristics. Following previous studies (Menzer and McFadden, 2017) we subdivide this data into low-density residential area (northern sectors, US-Minneapolis2) and irrigated grassland with few built structures (south, US-Minneapolis1). Each are given their own site timeseries and metadata, resulting in 21 datasets.

Table 7: Site wind sector exclusions. Sites with sensible and latent heat fluxes excluded because of land cover or land use differences by wind sectors as described in the reference provided. Maps of these sectors are provided in the site data collection (Lipson et al., 2022).

<table>
<thead>
<tr>
<th>Site name</th>
<th>Sectors excluded</th>
<th>Reason</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>FI-Kumpula</td>
<td>0-180°, 320-360°</td>
<td>surface inhomogeneity</td>
<td>(Karsisto et al., 2016)</td>
</tr>
<tr>
<td>FI-Torni</td>
<td>40-150°</td>
<td>flow interference from tower</td>
<td>(Järvi et al., 2018)</td>
</tr>
<tr>
<td>JP-Yoyogi</td>
<td>170-260°</td>
<td>surface inhomogeneity</td>
<td>(Ishidoya et al., 2020)</td>
</tr>
<tr>
<td>US-Minneapolis1</td>
<td>75-285°</td>
<td>surface inhomogeneity</td>
<td>(Menzer and McFadden, 2017)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>270-360°, 0-120°: surface inhomogeneity</td>
<td></td>
</tr>
</tbody>
</table>
5 Data availability

All site observational and characteristics data are openly available from https://doi.org/10.5281/zenodo.6590886 (Lipson et al., 2022) under a Creative Commons Attribution licence (CC-BY-4.0).

We recommend data users consult with site contributing authors and/or the coordination team early (i.e. planning stage) in projects that plan to use these data. Relevant contacts are included in site metadata.

6 Code Availability

Code used to process datasets are available at https://doi.org/10.5281/zenodo.6590942 (Lipson, M., 2022a).

Code used to create manuscript figures are available at https://doi.org/10.5281/zenodo.6590941 (Lipson, M., 2022b).

7 Acknowledgements

We would like to thank the vast number of people involved in day-to-day running of these sites that have been involved in instrument and tower installations (permitting, purchasing and site installation), routine (and unexpected event) maintenance, data collection, routine data processing and final data processing. We thank all those that have provided sites for the towers to be located and sometimes power and internet access. We acknowledge the essential funding for the instrumentation and other infrastructure, for staff (administrative, technical and scientific) and students for these activities. We also thank those who offered data for use in this project which are not included at this time.

The project coordinating team are supported by UNSW Sydney and the Australian Research Council (ARC) Centre of Excellence for Climate System Science (grant CE110001028), University of Reading, the Met Office and ERC urbisphere 855005. Computation support from the ARC Centre of Excellence for Climate Extremes (grant CE170100023) and National Computational Infrastructure (NCI) Australia. Contains modified Copernicus Climate Change Service Information. Site-affiliated acknowledgments are listed in Table 10.

Table 8: Funding acknowledgements for individual sites.

<table>
<thead>
<tr>
<th>Site</th>
<th>Contributing author</th>
<th>Site funding acknowledgements</th>
</tr>
</thead>
<tbody>
<tr>
<td>AU-Preston</td>
<td>Andrew Coutts, Nigel Tapper</td>
<td>-</td>
</tr>
<tr>
<td>AU-SurreyHills</td>
<td>Andrew Coutts, Nigel Tapper</td>
<td>-</td>
</tr>
<tr>
<td>CA-Sunset</td>
<td>Andreas Christen, Oliver Michels</td>
<td>Canadian Foundation for Climate and Atmospheric Sciences (CFCAS, Project &quot;Environmental Prediction in Canadian Cities (EPiCC)&quot;) and the Natural and Engineering Research Council of Canada (NSERC, RGPIN-03958, RGPAS-507854). Some instruments were supported by the Canada Foundation for Innovation (CFI, IF 2015, grant no. 33600) and BCKDF. We acknowledge the support of BC Hydro to operate the tower.</td>
</tr>
<tr>
<td>FI-Kumpula</td>
<td>Leena Järvi</td>
<td>ICOS Finland</td>
</tr>
</tbody>
</table>
### 7 Author Contributions

M.L., S.G. and M.B. conceived and coordinated the project, and prepared the protocols for contributing authors. M.L. collated site datasets, wrote processing code, developed and undertook analysis of bias correction methods, prepared figures, prepared datasets and drafted the manuscript with guidance from S.G. and M.B. All other authors (listed alphabetically) collected primary data, prepared site information, processed datasets for inclusion in the collection and contributed to the manuscript. Table 8 lists contributing author site affiliation.

The authors declare that they have no conflict of interest.
References


