3-15-2023

Development of a Benchmark Eddy Flux Evapotranspiration Dataset for Evaluation of Satellite-Driven Evapotranspiration Models Over the CONUS

John M. Volk
Desert Research Institute

Justin Huntington
Desert Research Institute

Forrest S. Melton
California State University, Monterey Bay

Richard Allen
University of Idaho

Martha C. Anderson
USDA-ARS Hydrology and Remote Sensing Laboratory

Recommended Citation

Volk, John M.; Huntington, Justin; Melton, Forrest S.; Allen, Richard; Anderson, Martha C.; Fisher, Joshua B.; Kilic, Ayse; Senay, Gabriel; Halverson, Gregory; Knipper, Kyle; Minor, Blake; Pearson, Christopher; Wang, Tianxin; Yang, Yun; Evett, Steven; French, Andrew N.; Jasoni, Richard; and Kustas, William, "Development of a Benchmark Eddy Flux Evapotranspiration Dataset for Evaluation of Satellite-Driven Evapotranspiration Models Over the CONUS" (2023). *AES Faculty Publications and Presentations*. 18.
https://digitalcommons.csumb.edu/aes_fac/18

Authors

John M. Volk, Justin Huntington, Forrest S. Melton, Richard Allen, Martha C. Anderson, Joshua B. Fisher, Ayse Kilic, Gabriel Senay, Gregory Halverson, Kyle Knipper, Blake Minor, Christopher Pearson, Tianxin Wang, Yun Yang, Steven Evett, Andrew N. French, Richard Jasoni, and William Kustas

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

journal homepage: www.elsevier.com/locate/agrformet

Development of a Benchmark Eddy Flux Evapotranspiration Dataset for Evaluation of Satellite-Driven Evapotranspiration Models Over the CONUS

John M. Volk [a,*], Justin Huntington [a], Forrest S. Melton [b,c], Richard Allen [d,e], Martha C. Anderson [f], Joshua B. Fisher [g], Ayse Kilic [e], Gabriel Senay [h], Gregory Halverson [i], Kyle Knipper [j], Blake Minor [a], Christopher Pearson [a], Tianxin Wang [k], Yun Yang [n], Steven Evett [l], Andrew N. French [m], Richard Jasoni [a], William Kustas [f]

[a] *Desert Research Institute, 2215 Raggio Pkwy, Reno, NV 89512, USA*
[b] *NASA Ames Research Center, Mail Stop 245-1, Moffett Field, CA 94035-1000, USA*
[c] *California State University, Monterey Bay, Seaside, CA 93955, USA*
[d] *University of Idaho, Moscow, ID 83844, USA*
[e] *University of Nebraska-Lincoln, Lincoln, NE 68588, USA*
[f] *USDA-ARS Hydrology and Remote Sensing Laboratory, Bldg. 007, Rm. 104, BARC-West, Beltsville, MD 20705-2350, USA*
[g] *University of California Los Angeles, Los Angeles, CA 90095, USA*
[h] *U.S. Geological Survey Earth Resources Observation & Science (EROS) Center/North Central Climate Adaptation Science Center and Faculty Affiliate with Ecosystem Science and Sustainability Colorado State University, Fort Collins, CO 80523-1499, USA*
[i] *NASA Jet Propulsion Laboratory, 4800 Oak Grove Drive M/S 233-300 Pasadena, CA 91109, USA*
[j] *USDA-ARS, Sustainable Agricultural Water Systems Unit, 239 Hopkins Road, Davis, CA 95616, USA*
[k] *University of California, Berkeley, Berkeley, CA 94720, USA*
[l] *USDA-ARS Conservation & Production Research Laboratory, 300 Simmons Road, Bushland, TX 79012, USA*
[m] *USDA-ARS US Arid-Land Agricultural Research Center, 21881 North Cardon Lane, Maricopa, AZ 85138, USA*
[n] *Mississippi State University, Mississippi State MS, 39759 USA*

ARTICLE INFO

Keywords:
Eddy covariance data
Evapotranspiration
Daily energy balance closure
Flux footprint prediction
Post-processing techniques

ABSTRACT

A large sample of ground-based evapotranspiration (ET) measurements made in the United States, primarily from eddy covariance systems, were post-processed to produce a benchmark ET dataset. The dataset was produced primarily to support the intercomparison and evaluation of the OpenET satellite-based remote sensing ET (RSET) models and could also be used to evaluate ET data from other models and approaches. OpenET is a web-based service that makes field-delineated and pixel-level ET estimates from well-established RSET models readily available to water managers, agricultural producers, and the public. The benchmark dataset is composed of flux and meteorological data from a variety of providers covering native vegetation and agricultural settings. Flux footprint predictions were developed for each station and included static flux footprints developed based on average wind direction and speed, as well as dynamic hourly footprints that were generated with a physically based model of upwind source area. The two footprint prediction methods were rigorously compared to evaluate their relative spatial coverage. Data from all sources were post-processed in a consistent and reproducible manner including data handling, gap-filling, temporal aggregation, and energy balance closure correction. The resulting dataset included 243,048 daily and 5,284 monthly ET values from 194 stations, with all data falling between 1995 and 2021. We assessed average daily energy imbalance using 172 flux sites with a total of 193,021 days of data, finding that overall turbulent fluxes were understated by about 12% on average relative to available energy. Multiple linear regression analyses indicated that daily average latent energy flux may be typically understated slightly more than sensible heat flux. This dataset was developed to provide a consistent reference to support evaluation of RSET data being developed for a wide range of applications related to water accounting and water resources management at field to watershed scales.

* Corresponding author.
E-mail address: John.Volk@dri.edu (J.M. Volk).

https://doi.org/10.1016/j.agrformet.2023.109307
Received 17 August 2022; Received in revised form 16 December 2022; Accepted 1 January 2023
Available online 27 January 2023

1. Introduction

Large sample datasets of ground-based evapotranspiration (ET) measurements from eddy covariance (EC) systems are highly sought after and are highly regarded by multiple disciplines in the Earth sciences as well as by the natural resource policy and management communities (Fisher et al., 2017). Such data have limited availability, due in part to the cost and specialized skills required to install and maintain EC systems and process the data. Therefore, it is common to use existing datasets and perform post-processing steps to ensure consistency across different datasets (e.g., Fisher et al., 2008). Post-processing decisions involving EC data aggregation, quality assurance and quality control (QA/QC), and energy balance closure corrections are frequently performed in an ad hoc manner within a study-limited scope rather than in a standardized or reproducible manner (Bayat et al., 2021). Although the use of existing EC measurements may be inappropriate in some circumstances due to system instrumentation, spatio-temporal limitations, or data processing performed, large sample ground-based ET datasets having reproducible and well documented data provenance are needed, particularly for benchmarking purposes (Pastorello et al., 2020). Because EC data continue to be collected and more EC stations continue to be installed, robust tools for ingesting new data records into existing datasets are also important.

Data integration across sites and networks results in data uncertainty due to site operation considerations and regional differences in topography and meteorological conditions (Allen et al., 2011). For example, systematic error from surface energy imbalance can arise due to station siting or techniques for processing the raw high frequency data (Moore, 1986; Finnigan et al., 2003); improper calibration, placement, and limitations of instrumentation (Kristensen et al., 1997; Högström and Smedman, 2004); and outlier detection and removal, gap-filling techniques, and other factors (e.g., Massman and Lee 2002; Fisher et al., 2007; Foken, 2008). A standardized post-processing and QA/QC routine cannot completely remove uncertainty and bias but can minimize them and ensure that results are reproducible (Moncrieff et al., 2004). The ONEFlux data pipeline for processing EC data, which was developed for the FLUXNET2015 dataset (Pastorello et al., 2020), advanced the standardization of eddy flux post-processing techniques. The EC closure correction methods we describe here are patterned after the procedures of ONEFlux, but our overall data processing pipeline differs from ONEFlux to additionally include data outside of FLUXNET (Baldocchi et al., 2001), to facilitate comparisons with remotely sensed ET (Melton et al., 2021), and to perform other flux data analyses such as energy balance closure assessments and footprint estimation. This led us to develop a suite of open-source Python tools for the reproducibility of methods and data provenance (e.g., Volk et al., 2021) that complement and add to the functionality provided by the ONEFlux tool.

We identified energy balance closure corrections to daily average latent energy fluxes as an important step for improved ET estimates, mainly because non-closure is widely reported and known to exist at EC towers, particularly those having heterogeneous land cover, tall canopy heights, non-flat terrain, and during periods of high stability/weak turbulence or strong advection (e.g. Aubinet et al., 2000; Twine et al., 2000; Wilson et al., 2002; Barr et al., 2006; Fisher et al., 2007; Foken, 2008; Mauder et al., 2013). Conducting closure corrections on daily aggregated data as opposed to hourly or higher frequency data was chosen for reasons associated with the effects of diurnal hysteresis in fluxes and phase lags between energy balance components and the environmental variables that affect them (e.g., Li et al., 2008; Gao et al., 2010; Lin et al., 2019; Dhungel et al., 2021). For example, Leuning et al. (2012) suggest that temporal variations in energy storage (in soil, air, and biomass), which tend to reduce at the daily scale, can result in an additional 15% of understatement of turbulent fluxes relative to radiative fluxes at hourly timescales. The sources and patterns of energy balance closure error remain a key area of study in the scientific community (e.g., Bambach et al., 2022 and Dhungel et al., 2021); therefore, we leveraged data from a large sample of flux sites (251), ingested to form the benchmark ET dataset, to analyze energy balance closure error and its relation to land cover type and seasonality. The dataset is also used to evaluate the relative magnitude of closure error expressed between latent and sensible heat fluxes. In addition, assessment at daily rather than half-hourly timesteps, as reported by most large sample studies on EC closure error, presents a unique insight into the energy balance closure problem (Stoy et al., 2013; Wilson et al., 2002; Twine et al., 2000; Foken, 2008).

A key step for comparing remotely sensed ET (RSET) against *in situ* flux measurements is the development of accurate and representative flux footprints for sampling of pixels in the source area of flux towers. Several footprint approaches are available for RSET sampling ranging from simple buffers centered on flux tower locations (e.g., Fisher et al., 2020) to physically based, temporally dynamic footprint models that consider atmospheric conditions such as aerodynamic roughness, stability, wind direction and speed, and measurement height (Kljun et al., 2015; Kormann and Meixner, 2001). In sites with high levels of surface heterogeneity, the accuracy of footprint sampling techniques become more important (Chu et al., 2021); however, large-scale comparisons between footprint methods are rare. Using the benchmark dataset, we compare the relative utility of simple fixed pixel-grid footprints and dynamic footprints generated using the Kljun et al. (2015) model for routine evaluation of Landsat-scale RSET data.

Our effort to develop a large sample benchmark ET dataset, flux station footprints, and ancillary tools is primarily intended for the multiphase intercomparison of satellite-based RSET models that are part of the OpenET initiative (Melton et al., 2021). OpenET is a web-based platform that uses an ensemble of satellite-driven ET models and leverages Google Cloud and Earth Engine services to make RSET data publicly accessible via a web application and application programming interface (API). Six well established satellite-based RSET models are included in OpenET: DisALEXI, eeMETRIC, geeSEBAL, PT-JPL, SIMS, and SSEBop (Melton et al., 2021). OpenET models are primarily based on data from the Landsat Thermal Infrared Sensor (TIRS) and the Operational Land Imager (OLI), and provide ET data at daily, monthly, and annual time scales at the Landsat spatial resolution of 30 × 30 m. OpenET also provides pre-computed data time series for millions of individual fields, making it a powerful tool for understanding agricultural water use and ET across the land surface. We emphasize, however, that the flux dataset presented here can be used for a variety of other applications that require accurate estimates of *in situ* ET such as hydrologic and land surface modeling, carbon and energy cycling, and ecological studies.

The remainder of this paper contains a description of flux data sources, data ingestion procedures, QA/QC and closure correction methods, and flux footprint development methods. After the method descriptions, we discuss energy balance closure results from the EC towers, coherency between temporally dynamic and statically defined flux footprint predictions, and dataset limitations and future directions. This paper is accompanied by a short 'Data in Brief' article (Volk et al., 2022) that includes additional technical details not covered here regarding the flux data, such as file formatting and structure, meteorological calculations, and diagnostic graphics. The accompanying paper also includes a link to a public data repository with an archived version of the post-processed benchmark flux dataset, which consists of daily and monthly aggregated flux and ET data and diagnostic graphics for each site.

1.1. *Data sources*

The OpenET benchmark dataset began with the collection of data from 328 EC sites and ET data from an additional 4 weighing lysimeter sites, 11 Bowen ratio sites, and 3 residual energy balance (REB) instrumented sites. The analysis presented in this paper is focused on EC station results. Additional background on the Bowen ratio and lysimeter instrumented sites is provided in the supplementary material (Text S1 and S2). Data providers include multiple organizations and teams; however, most data (260 sites) were retrieved from EC sites within the AmeriFlux network (Pastorello et al., 2020). EC stations provide the most direct and generally accurate measurements of latent and sensible heat flux that are available at scale over many locations, vegetation, and climate conditions, and many are maintained for extended periods of time (Baldocchi et al., 2001; Baldocchi, 2014) allowing for robust intercomparisons in time and space.

The ET stations included in this study are located in a variety of land use and land cover types including irrigated and non-irrigated agriculture, as well as a variety of non-agricultural vegetation sites. The stations are well distributed over the contiguous United States (CONUS) with higher station density in areas of intensive water use research, e.g., in the Central Valley of California, shrublands in Nevada, and the Corn Belt region of the midwestern United States (Fig. 1; Supplementary Table S6). In addition to AmeriFlux, EC data providers include: the United States Department of Agriculture Agricultural Research Service (USDA ARS) (Text S3), including four sites from the GRAPEX program (Text S4) (Kustas et al., 2018), U.S. Geological Survey Nevada Water Science Center ET studies (https://nevada.usgs.gov/et/), California State University Monterey Bay, Desert Research Institute (Text S5), Texas A&M University, and the Delta-Flux network (Runkle et al., 2017). Four weighing lysimeter sites (Text S2) deployed in agricultural fields near Bushland Texas and operated by USDA ARS (Evett et al., 2016), and 8 Bowen ratio sites mostly in native phreatophyte shrublands and grasslands in Nevada (Text S1) operated by the USGS were also included in the final benchmark dataset. Additional ET measurements from non-EC systems were included because they provide additional spatial-temporal coverage to the dataset. For example, the lysimeter sites in Texas are the only measurements we have for annual crops in the large state. More information on data sources including station principal investigators and team members, acknowledgments, land cover, and other metadata is listed in Supplementary Table S6 and the accompanying data article (Volk et al., 2022).

General land cover classifications are provided for AmeriFlux sites using the International Geosphere-Biosphere Programme (IGBP) scheme; however, we reclassified some sites based on a detailed inspection of metadata, aerial and satellite imagery, and literature review. For example, a few sites originally classified as grasslands were reclassified as croplands after reading Principal Investigator (PI) notes that indicated any irrigation and/or harvesting of the vegetation. Information on the AmeriFlux sites that had land cover type reclassification are listed in Text S5. We also lumped deciduous forested sites with mixed forests, and further classified cropland sites into annual crops, vegetable crops, orchards, and vineyards. Overall, using our classifications, the final benchmark dataset includes 75 cropland sites (59 annual crops, 7 vegetable crops, 5 orchards, and 4 vineyards), 33 grassland sites, 36 shrubland sites, 18 mixed forest sites, 23 conifer sites, and 9 wetland or riparian sites, for a total of 194 station locations. The initial pool of over 300 stations was reduced because of data availability and energy balance closure requirements, which are described in detail in the following sections.

To assess and build confidence in the EC data, we conducted energy balance closure analyses at the daily timestep. Measurements of energy balance components included: latent energy (LE), sensible heat flux (H), net radiation (R_n), and soil heat flux (G), all of which are required for energy balance closure analysis and correction. Of the 328 EC sites ingested with LE measurements, 256 also had records of H, R_n, and G at the daily timescale after limited gap-filling. If a site was missing one or more of the four primary energy balance components throughout the site data record, we could not assess the energy balance closure, and the site was excluded from the final dataset. For all but a handful of Bowen ratio instrumented sites where daily ET estimates were provided, input data temporal frequency was half-hourly, which is the common averaging period of most EC data processing software, e.g., LI-COR EddyPro (LI-COR Biosciences, Lincoln, Nebraska).

1.2. Flux data post-processing: ingestion, filtering and gap-filling, closure corrections, time integration, and QA/QC

Post-processing and QA/QC procedures for EC flux data are outlined by the following steps: (1) gap-filling of missing or faulty half-hourly or hourly energy balance components; (2) daily aggregation; (3) energy balance closure correction; (4) gap-filling of daily ET to produce a complete record; (5) monthly aggregation; and (6) visual inspection and

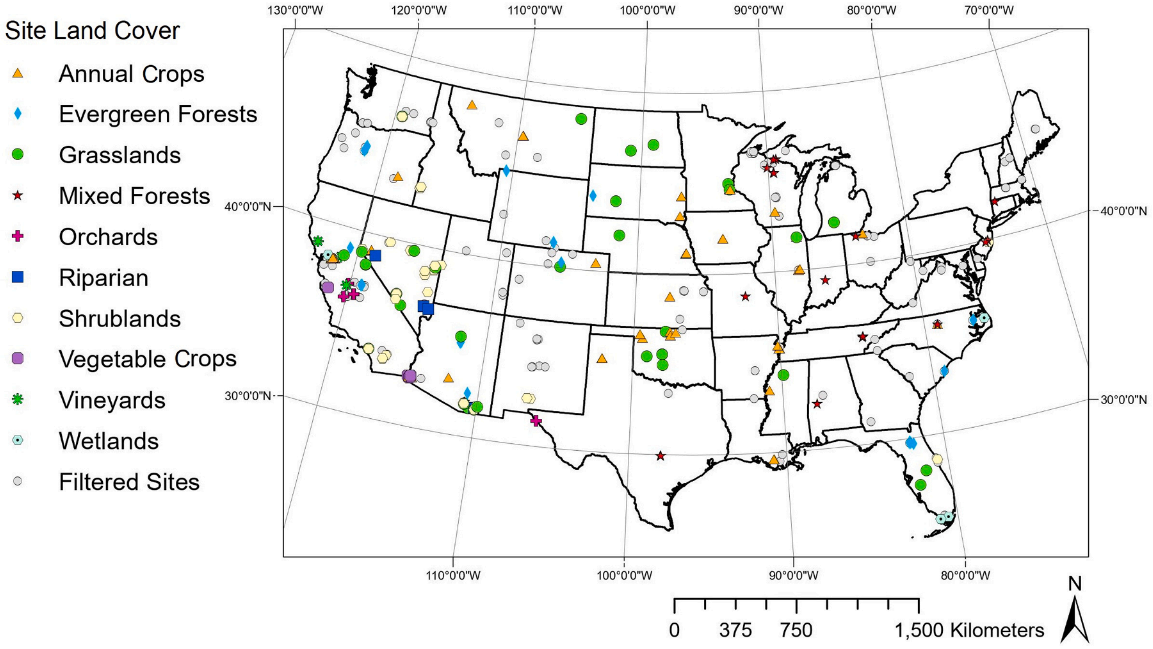

Fig. 1. Locations of EC and other ground-based stations used in the OpenET benchmark dataset. Filtered sites are those that were ingested and have daily ET but were not included in the benchmark dataset due to energy balance closure issues or data availability.

screening of post-processed data. These were the same steps followed in the recent ECOSTRESS ET validation study (Fisher et al., 2020). Explanations of these procedures are given in the following section with additional technical details in Volk et al. (2022). An open-source Python package "flux-data-qaqc" was developed to perform reproducible procedures for steps 1-5 and to produce visual tools for step 6 (Volk et al., 2021). The software is hosted on GitHub, PyPI, and has online documentation including a comprehensive user tutorial. The automated procedures result in daily and monthly ET time series that were generated in a standardized manner. An automated software approach also facilitates future integration of new data records into the benchmark ET dataset.

1.2.1. Data ingestion

AmeriFlux EC data were downloaded from the AmeriFlux online data archive (https://ameriflux.lbl.gov/), and EC station data for other sites were retrieved from site PIs and their teams. Because AmeriFlux stations often record multiple sensors for measurement of a variable (e.g., multiple soil heat flux plates to sample spatial heterogeneity), we used a standardized algorithm for selecting a preferred record and/or for averaging multiple records into a single record using standard AmeriFlux variable flags. The algorithm is provided in Text S8, and it gives priority to PI-approved, gap-filled, and aggregated records. When no preferred records exist, the algorithm falls back on using the average from multiple sensors. For EC data from other non-AmeriFlux providers, data reduction from multiple sensors was based on the recommendations of site PIs and their teams.

In addition to energy balance data, tower meteorological measurements for variables such as air temperature and vapor pressure were ingested when available. Other variables, such as saturation vapor pressure and potential solar radiation, were estimated from initial half hourly data following methods put forth by the American Society of Civil Engineers (ASCE) (Allen et al., 2005). Daily ASCE standardized Penman-Monteith grass reference ET was calculated (Allen et al., 2005) at sites having sufficient data, and daily gridMET precipitation and grass and alfalfa reference ET data (Abatzoglou, 2013) were downloaded for all sites. Volk et al. (2022) provides a full list of ingested meteorological variables, derived meteorological variables, and their calculation.

Site meteorological data were also used for flux footprint estimation and flux data QA/QC. For example, air temperature was used to correct the latent heat of vaporization, and wind direction and speed were used for the generation of wind rose diagrams for footprint generation and validation. For flux data QA/QC, meteorological data were used as a reference for site conditions and for visual-based assessment of final ET estimates on an individual, site-to-site basis. Meteorological and flux data were archived as time series and interactive graphical files for each site at daily and monthly timesteps, and derived meteorological data were calculated using the "flux-data-qaqc" Python package version 0.1.6 (Volk et al., 2021; Volk et al., 2022).

1.2.2. Initial filtering, gap-filling, and computation of 24hr-average fluxes

Initial gap-filling of half-hourly and hourly LE, H, R_n, and G [W m^{-2}] followed a simple method: gaps up to 4 hours long during the night (defined as periods with $R_n < 0$) and 2 hours during the day (defined as $R_n >= 0$) were linearly interpolated. We limited gaps to two consecutive hours during daytime as a conservative measure. If gaps still existed after interpolation, e.g., a daytime gap that was longer than 2 hours, then the daily flux value was flagged as a gap. The total number of sub-daily gaps that were interpolated per day were recorded to allow for post-filtering of days with excessive gaps; for example, days that had multiple short gaps. The resulting days that were not flagged as gaps after this procedure were averaged to daily flux/energy components [W m^{-2}].

1.2.3. Energy balance closure correction and ET calculation

Several methods have been used in published studies to enforce energy balance closure in EC flux datasets, and to adjust LE and H such that $R_n - G = H + LE$ at some timescale, typically half-hourly. The most commonly used methods include (1) residual closure, where the total energy imbalance is assigned to the latent energy term (e.g., Prueger and Kustas, 2005); (2) Bowen ratio closure, where LE and H are both adjusted while preserving the observed Bowen ratio (H/LE) (Twine et al., 2000); and, (3) energy balance ratio (EBR) closure, where $EBR = (LE + H)/(R_n - G)$, and in which LE and H are both adjusted such that the EBR averages to 1 over some timescale (Pastorello et al., 2020). Although each approach has advantages and disadvantages, here we adopt the EBR closure technique for multiple reasons. First, this choice is consistent with the methods used in the FLUXNET2015/ONEFlux dataset generation (Pastorello et al., 2020). Compared with the residual method, EBR yields more conservative corrections to LE, particularly during periods of low LE flux where residual corrections can be unreasonably large. Also, because the prescribed EBR method does not force closure at a daily time step and instead uses filtered EBR over sliding windows (e.g., 15 days), the resulting correction factors are less influenced by short-term anomalies in the EBR. Like most methods, this approach has associated trade-offs. If local extrema in the EBR window are accurate, this method may provide a correction that is too conservative on the date of the extrema; on the other hand, if locally extreme EBR values are not dependable or realistic, then the sliding window EBR technique method will dampen them and move the correction factor towards the typical observed values around a given date.

The specific steps we used for energy balance closure correction on daily fluxes are as follows, with the superscript "#" signifying a step included here and not used in the ONEFlux processing pipeline. The additional steps are checks for extreme daily corrected LE values that fall outside of what are physically reasonable. Extreme LE values can arise due to anomalous EBR values resulting from faulty initial data or during periods when fluxes are near zero.

(1) First, EBR is calculated using daily averages of flux components, and days having EBR values outside of 1.5 times the interquartile range are removed to limit skewing of EBR-based adjustments by extreme values.
(2) For each day, the median EBR is selected from a centered 15-day sliding window. If there are less than 11 days in the window to determine a median value, then the average is used from a centered 11-day window. #If the absolute value of the EBR reciprocal, |1/EBR|, is greater than or equal to 2, or less than or equal to 0.5, or if LE times the EBR reciprocal (LE/EBR) is greater than 800 or less than -100 [W m^{-2}], then those daily EBR values in the window are left as gaps.
(3) If step 2 fails (i.e., the EBR could not be calculated within the 11-day window), then compute the EBR climatology or the average from each day of year on record and apply an 11-day centered moving average to extract an EBR value for each gap day. #Apply the same check for extreme EBR values as shown after step 2.
(4) Correct daily LE and H by multiplying them by the reciprocal of the daily filtered EBR as produced by steps 1-3 (Fig. 2).

For graphical illustration of these steps, please see the "flux-data-qaqc" online documentation.

This closure correction technique rarely results in perfect energy balance on any given date (Fig. 2), but turbulent fluxes are adjusted such that closure converges to 1 over the sliding window periods, although less so during time periods of high variability in the EBR.

After the steps described above, we use daily average LE flux [W m^{-2}] to estimate the average ET rate [mm/day] with consideration of the effect of daily average air temperature on the latent heat of vaporization, following the method of Harrison (1963). The adjustment to the latent heat of vaporization caused by air temperature typically results in a slight increase in ET, up to about 1%, with slightly more of an effect when latent energy flux is relatively high.

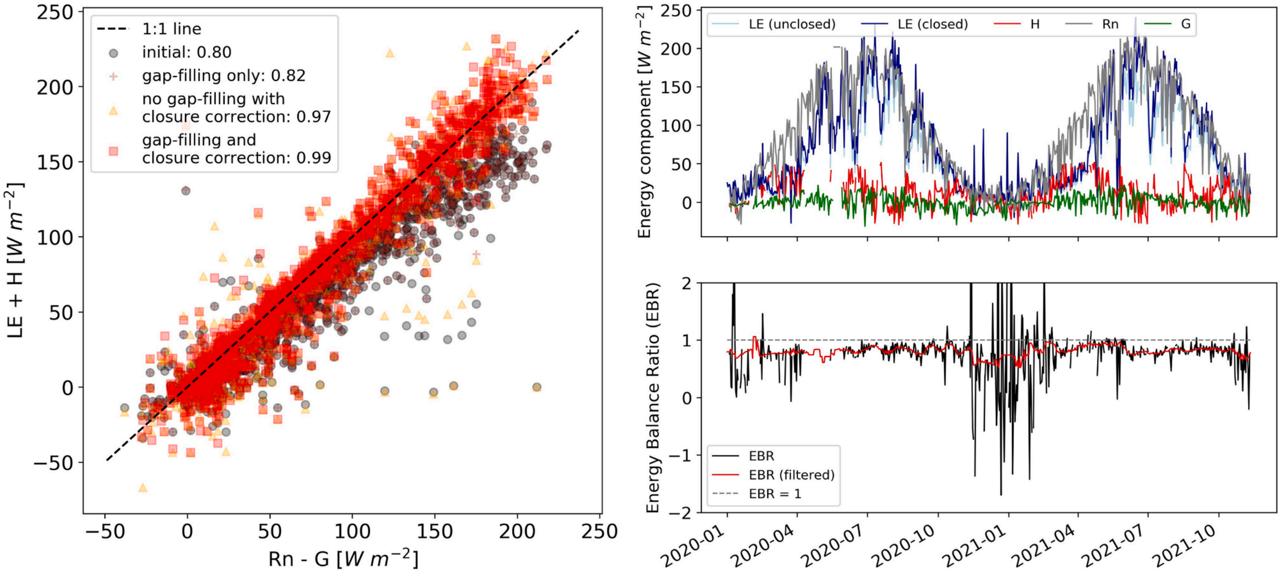

Fig. 2. Daily average turbulent fluxes and closure results before and after applying half-hourly gap-filling and energy balance ratio closure corrections (left) using data from an EC tower in an irrigated alfalfa field in the Harney Basin, Oregon. The average daily energy balance closure, as defined by the linear regression slope forced through the origin, is shown in the legend for each processing step. The upper right plot shows daily average energy balance components after gap-filling, including latent energy flux after closure correction. The lower right plot shows the daily energy balance ratio including the “filtered” energy balance ratio that is calculated using sliding windows and used to correct daily turbulent fluxes.

1.2.4. Gap-filling daily ET

After energy balance closure correction, gaps in daily ET were filled. We used the fraction of reference ET (EToF) over the gap period, using 4-km gridMET grass reference ET (ETo) (Abatzoglou, 2013) at the tower location as a scaling flux. The daily EToF was calculated as the ratio of the closed ET to gridMET ETo. The EToF time series was then filtered by removing values outside of 1.5 of the interquartile range, smoothed using a 15-day moving average, and then linearly interpolated to fill remaining gaps. Daily ET gaps were filled using the filtered EToF multiplied by ETo (Allen et al., 2007). From our observations, the EToF gap-filling method does well if gap lengths are not excessively long. Gap-filled daily ET values were calculated to facilitate computation of monthly total ET values at each site, but the gap-filled daily values were not included in the OpenET benchmark dataset, nor were monthly ET totals that include more than five gap-filled days.

1.2.5. Seasonal ET

For each flux site, prior to energy balance closure correction and daily gap-filling, average energy balance closure was assessed for growing versus non-growing seasons using daily average LE fluxes computed from half-hourly gap-filled data. Water usage and energy balance closure are often different in magnitude and variability during different seasons, and it is helpful to separate these periods for scientific and operational applications. Yearly growing season start, end, and length datasets from 1980-2020 were estimated for each study site using the full gridMET climate dataset based on cumulative growing degree day (CGDD) and killing frost temperature thresholds (Abatzoglou, 2013) (Fig. 3). CGDD based on daily average temperature has shown to be a good metric for estimating plant available energy and phenology (e.g., Allen and Robison, 2007; Huntington and Allen, 2009; Sammis et al., 1985; Wright, 2001). Daily minimum temperature was used to identify

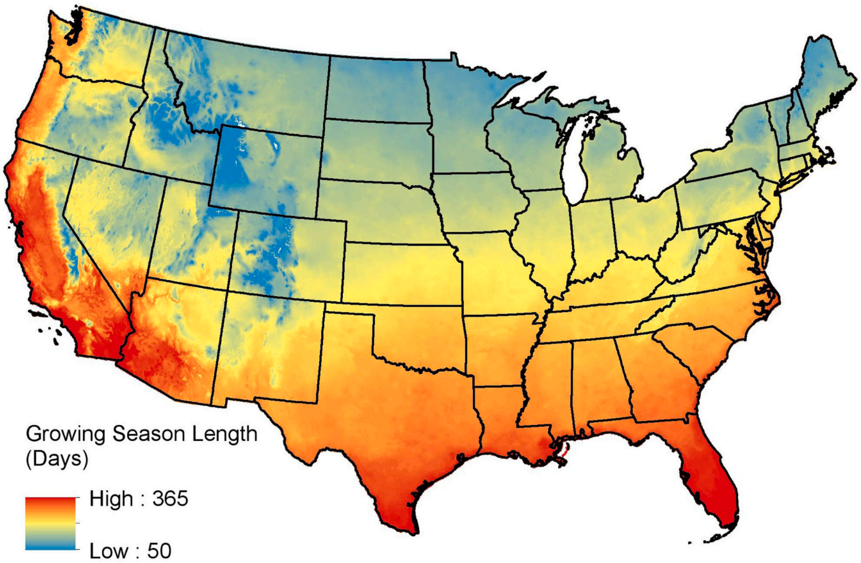

Fig. 3. Map of growing season length in days as determined from cumulative growing degree days and killing frost using data from 1980-2020.

the first occurrence of killing frost temperatures, when most plants have reached a stage of full senescence and the growing season ends.

The CGDD map was computed as follows. Starting on January 1st of a given year, the daily average temperatures were collected. The start date of the annual growing season was defined as the day where the running sum of daily temperature reached the threshold of 300 °C, and the end of the season was defined as the first date where the daily minimum temperature was -2 °C or less. These temperature thresholds were based on general historical plant phenology and growing practices as well as comparisons with remotely sensed normalized difference vegetation data. A start date of January 1st was applied for years when the killing frost threshold was not reached during the previous year. Killing frost was assumed to not occur before DOY 200 to avoid false detection related to early season cold snaps. Data from 1979 were used as a spin up year to initialize estimates and start dates at locations with year-round growth. All data were processed and compiled using the Google Earth Engine analysis platform (Gorelick et al., 2017). Although application of single CGDD and killing frost thresholds cannot capture phenology for all vegetation types, dates established by this analysis are applied to distinguish between generally active and dormant time periods.

1.2.6. *Final data filtering and QA/QC*

A final screening of EC data was based on data availability, energy balance closure, and visual inspection. A subset of 194 of the 328 sites with ET data were selected to form the benchmark dataset collection for the OpenET intercomparison and accuracy assessment (Melton et al., 2021). To be included, data records must include all four energy balance components (LE, H, R_n, and G) such that energy imbalance can be assessed and corrections can be performed on daily averaged fluxes. For monthly data records, a maximum of five gap-filled days per month were allowed, as described above. Next the CGDD-based growing and non-growing season periods were used to calculate seasonal closure at each site, based on the slope of the least squares linear regression forced through the origin, i.e., intercept = 0, where $x = R_n - G$ and $y = LE + H$. Only sites with growing season closure > 0.75 and non-growing season closure > 0.6 were selected for the benchmark dataset. High energy imbalance, particularly during the growing season, indicates that the EC technique may be inappropriate at a given site, which may be due to a variety of factors including tower and instrument placement or large-scale exchanges that are not fully captured by the EC instrumentation (e.g., Foken, 2008).

Sites that passed the energy balance criteria, and the Bowen Ratio and lysimeter sites, underwent visual QA/QC review to identify data quality concerns not captured by gap-filling and closure correction. For example, flat lines, repeating patterns, or other systematic and obvious data artifacts that may be caused by instrumentation error were concerns; these data were flagged, and faulty data were removed. Although the energy balance closure correction filters outlier ET values, it depends on the daily filtered EBR from sliding windows, and therefore it does not consider all sources of potential error in daily ET. This is one reason that the manual QA/QC is necessary. ET values were also checked against corresponding energy balance values and meteorological data, and if values were anomalous over brief time periods, the half-hourly fluxes were inspected and faulty data were removed, and the gap-filling and closure correction routines were rerun. In rare cases, sites were removed entirely from the benchmark dataset based on qualitative decisions. For example, all three residual energy balance sites that were evaluated had sparse and consistently questionable ET magnitudes, so these sites were excluded. In addition to energy imbalance and meteorological data, visual-based data filtering was informed by gridded ETo (Abatzoglou, 2013) and EToF as well as looking at the tower location and aerial images for obvious obstructions and land cover issues that may affect the turbulence at the site in a way that violates the site requirements for the EC technique.

The resulting 194 sites that passed our gap-fill and closure criteria and that underwent visual QA/QC and filtering comprise the benchmark ET dataset for OpenET (Melton et al., 2021). The benchmark dataset contains a total of 243,048 days (about 665 and a half years) and 5,284 months (about 440 and a half years) of ET that have been corrected for energy imbalance, and most sites data records fall within the last two decades.

1.2.7. *Data availability*

Post-processed flux and meteorological data, as well as diagnostic plots, for 161 stations that have been used for the OpenET second phase intercomparison and accuracy assessment will be archived and made public (Volk et al., 2022). The remaining 33 sites are being held back for use in future blind model evaluations, and their data will be published at that time; however, they are included in the energy balance closure and flux footprint analysis here. The held-back sites were chosen randomly from the subset of sites that have not been previously used for ground validation with any of the OpenET models.

1.3. *Multiple linear regression of energy balance components*

To investigate the validity of the EBR closure correction and to better understand the relations between daily energy balance closure and individual daily average turbulent fluxes, multiple least squares linear regression (MLR) analysis was performed. The analysis identifies systematic under- or over- estimation of turbulent fluxes, LE and H, which are often assumed to have similar energy balance error using the EC technique (e.g., Pastorello et al., 2020; Twine et al., 2000). The results will also be useful for assessing data quality; for example, if some sites have anomalous regression coefficients for G, then the soil heat flux data may have quality issues. This approach depends on the accuracy of G and R_n measurements, which are the dependent variables. Using available energy (R_n – G) as the dependent variable:

$$R_n - G = c_0LE + c_1H \tag{1}$$

and, assuming only net radiation measurements are reliable:

$$R_n = c_0LE + c_1H + c_2G \tag{2}$$

This analysis leveraged daily gap-filled records of energy balance components for 172 EC sites that passed initial screening based on closure and qualitative review and was limited to those sites with at least 30 days of data. The regression analysis was performed for each site using the “scikit-learn” Python module, version 0.22 (Pedregosa et al., 2011). The accuracy of the regression at each site was evaluated using the coefficient of determination, calculated as the square of the Pearson correlation coefficient, and the root-mean-square-error.

1.4. *Flux footprint predictions*

Flux footprints were produced for each ET site to accurately sample RSET pixels for comparison with observations. Two types of footprints were developed: (1) simple static gridded footprints of Landsat pixels (e.g., 3 × 3); and (2) the Kljun et al. (2015) 2-dimensional flux footprint model was used to create daily and monthly footprints weighted by hourly ETo. Static footprints were based on the long-term wind direction and speed, whereas the flux footprint model was based on hourly wind dynamics, surface roughness estimates, and atmospheric stability measurements. The footprints (static and dynamic) were developed at the spatial resolution of Landsat (30 × 30 m) and use the Landsat geographic projection for direct comparisons of pixel ET estimates from OpenET RSET models, which calculate ET data using the Landsat spatial reference system and resolution.

1.4.1. *Static footprints*

Satellite pixel grids were generated following an approach similar to Fisher et al. (2020), who applied 3 × 3 and 5 × 5 (70-m pixel resolution) grids centered on flux towers. However, instead of using a fixed grid centered on the tower, we selected more optimal upwind grid locations based on the wind rose from 6:00 to 18:00 local time (Fig. 4b).

A schematic processing diagram is provided in Fig. 4a. For each flux tower, a 500-m buffer is created, and the composited Landsat image is clipped into the buffer size. Then, the clipped image is used to create the sample grid (510-m by 510-m). Lastly, different sets of pixel grids are selected with the constraint of the wind rose to represent the areal footprint. For each site, 3 × 3 and 5 × 5 or 7 × 7 grids were produced, depending on the field size, geometry, and surroundings. For example, the 7 × 7 grid was typically preferred, but the 5 × 5 grid may be selected to prevent inclusion of confounding features if the 7 × 7 grid includes open water, roads, or other non-agricultural areas adjacent to the flux station location. Additionally, if the pixel containing the flux tower contains features that are not representative (e.g., concrete), the grids are shifted to capture the area that is representative for the site.

1.4.2. *Dynamic ETo weighted footprints*

Wind direction and speed, as well as turbulence structure, are temporally dynamic, and flux source areas are as well. The Kljun et al. (2015) 2-dimensional flux footprint parametric model provides an efficient and practical method to estimate areal extent and location of flux source area. The method allows for variable measurement heights, accounts for average surface roughness, and includes crosswind distributions (upwind footprint width), all of which are important for sampling

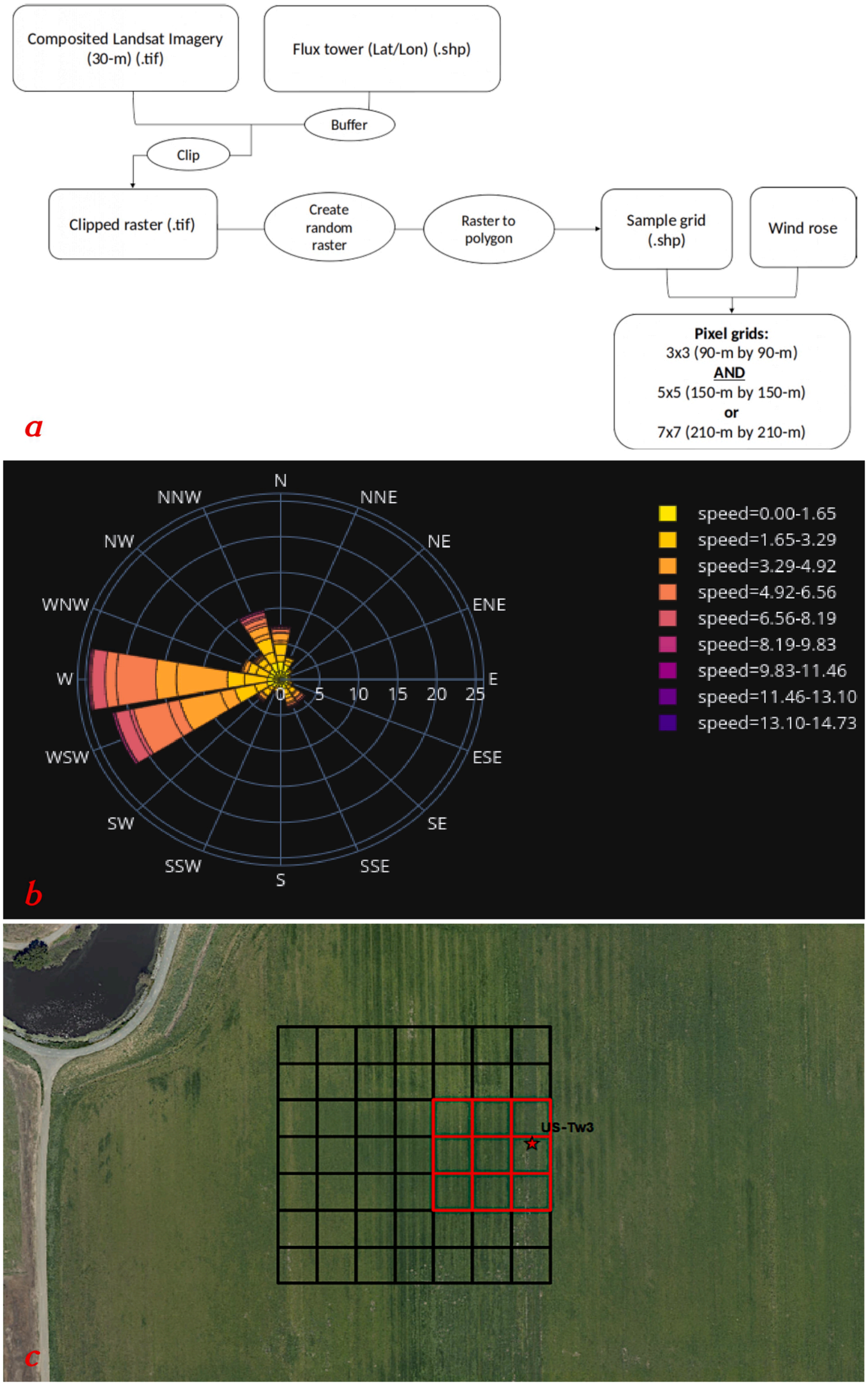

Fig. 4. (a) Schematic showing processing steps for generating the Landsat pixel grid static flux footprints; (b) Long-term daytime wind rose generated for AmeriFlux EC site US-Tw3; (c) 3 × 3 and 7 × 7 footprint grids for site US-Tw3. We credit the National Agriculture Imagery Program (NAIP) for the basemap used in Fig. 4c.

of RSET pixels. The Kljun et al. (2015) algorithm uses a scaling approach to estimate the crosswind distribution of the footprint that acts to improve numerical efficiency. Footprints were computed at hourly timesteps, and then averaged to daily and monthly footprints using an ETo-based weighted averaging approach using gridded hourly ETo calculated from the North American Land Data Assimilation System version 2 (NLDAS2) data and extracted at the ET stations (Xia et al., 2012). Hourly weighting by ETo gives priority to the footprint predictions during periods of relatively high evaporative demand. The weighting approach was based on a method developed by researchers at Utah State University (Richard Allen and David Eckhardt, written communication Aug. 19, 2020) with slight modifications. Full processing steps of the temporally dynamic footprint production are as follows:

1. Collect or estimate hourly average input parameters; "*" signifies that the parameter is sometimes estimated using various methods described in the next section: *measurement height above zero-plane displacement height ($z_m - d$); *Monin-Obukhov length (L); *friction velocity (u_*); *aerodynamic roughness length (z_0); horizontal wind speed (u); wind direction; *standard deviation of the crosswind component of wind due to turbulence (σ_v); and *planetary boundary layer height.
2. Apply the Kljun et al. (2015) 2-dimensional flux footprint model (Python version) on hourly data centered on tower locations using 30-m resolution and a 600 m square extent. Use daytime hours from 6:00 to 20:00 local time in the footprint model. Normalize each hourly footprint so that the sum of pixels equals one.
3. Use hourly NLDAS2 data to calculate hourly ETo. Specifically, vapor pressure is estimated from air pressure and specific humidity, solar radiation, wind speed, and average air temperatures are downloaded as GRIB binary files from the NASA Goddard Earth Sciences Data and Information Services Center online application (https://disc.gsfc.nasa.gov/) and point data are extracted. The NLDAS2 data are used to calculate the (ASCE) Penman-Monteith hourly standardized grass reference ET (Allen et al., 2005), again for daytime hours. For each hour, calculate the fraction of ETo to the daily total ETo.
4. Use the hourly ETo fractions from step 3 to scale the corresponding hourly footprints of step 2, then sum the weighted hourly footprint images over the day to create daily weighted footprints, requiring a minimum of 5 hours of data per day for daily footprints. Normalize the final daily footprint such that its pixel values sum to 1. For monthly footprints, use the hourly ETo fractions to normalize them so that they sum to 1 using the total valid hours of footprint data within the month; a minimum of 260 daytime hours are required to generate a monthly footprint.
5. Check that each daily and monthly weighted footprint sum to 1.
6. Output daily and monthly footprints as georeferenced rasters using their local UTM (Universal Transverse Mercator) zone and geotransform them such that they align with Landsat pixels and can be used to efficiently extract RSET model output.

All steps involved in this footprint production process (listed above) have been made reproducible via open-source Python scripts hosted in the "flux-data-footprint" GitHub repository, which includes examples and documentation.

1.4.3. Parameter estimation

Input parameter estimation for the Kljun et al. (2015) footprint model was often required due to limited wind and turbulence data availability at EC sites. Beforehand, site PIs were contacted regarding missing data. For sites where the height of the 3-dimensional anemometer was not reported and no imagery was available, we assumed the anemometer was near the top of the tower at the highest measurement height listed in the tower metadata. Most commonly σ_v, or the standard deviation of lateral wind velocity, was the only missing input parameter. For sites where hourly u_* measurements were available, σ_v was estimated as $\sigma_v = 1.9 * u_*$ based on a literature review (e.g., Smedman 1988; Pahlow et al., 2001; Vickers and Mahrt, 2007) and empirical evidence from linear regression results on all sites in the dataset (see Table S9). The average least squares regression coefficient for u_* from 47 EC sites was 1.9; however, the relation had substantial scatter between hourlyσ_v and u_* observed at certain sites, which indicates a linear model is not robust, and additional information such as canopy height and roughness would improve the parameterization.

To estimate d and z_0, hourly canopy height (h_c) was first estimated (if unknown) using the method described in Pennypacker and Baldocchi (2016),

$$h_c = \frac{z_m}{\frac{d}{h_c} + \frac{z_0 e^{ku/u_*}}{h_c}} \tag{3}$$

where d/h_c is assumed to have a value of 0.6, and z_0/h_c is assumed to have a value of 0.1, k is the von Kármán constant = 0.4, and u is the horizontal wind speed at height z_m. We note that the assumed values for d/h_c and z_0/h_c, as used by Pennypacker and Baldocchi (2016), are functions of leaf area index and could be improved, particularly in forested sites where h_c estimation is sensitive to these values. The Pennypacker and Baldocchi (2016) method has been shown to be accurate across a variety of land cover types during stable conditions (Chu et al., 2018). As a conservative measure, we only applied the canopy height estimation on hourly data that passed the strict stability requirements of $|z_m/L| < 0.03$ as suggested by Pennypacker and Baldocchi (2016) and $u_* > 0.2$ m/s (Papale et al., 2006), which signifies appropriate turbulence conditions for the eddy covariance technique. We applied additional filtering and gap filling to the estimated h_c data to produce full hourly data records. This procedure involved: (1) creation of a time series of the original hourly canopy height using a centered 720 hour (30 day) window using an exponentially weighted window with tau = 5; (2) filling remaining gaps with the long-term average canopy height from step 1; and (3) smoothing the result using a 720 hour (30 day) centered, moving average. The rationale for the initial moving average with an exponential weighted window (as opposed to uniformly weighted) was to prevent skewing h_c by values that are several days or weeks out from a data point within the 30-day window. The resulting time series of h_c were reasonable estimates that are weighted towards canopy estimates during periods of near-neutrality and have smoother transitions of canopy height estimates in between such periods (Fig. S10).

Zero-plane displacement height was estimated based on land cover type and h_c, for forests:

$$d = (2/3)^* h_c \tag{4}$$

(Stull, 1988; Arya, 1998); for cropland and grassland sites:

$$d = 10^{0.979^* log(h_c) - 0.154} \tag{5}$$

(Rosenberg et al., 1983); and for other surfaces:

$$d = (3/4)^* h_c \tag{6}$$

(Kaimal and Finnigan, 1994). Aerodynamic roughness length, when not provided, was estimated as $0.1^* h_c$. Given the broad range in vegetation type, canopy densities, and uncertainty in h_c estimates were truncated from the $0.12^* h_c$ as described in Jensen and Allen (2016). The Monin-Obukhov L when not provided was estimated as:

$$L = -\frac{\rho_{air} \times Cp \times T \times u^{*3}}{k^* g^* H} \tag{7}$$

where ρ_{air} is air density [kg m^{-3}], Cp is the specific heat of air at constant pressure [J kg^{-1} K^{-1}] = 1005, T is average air temperature [degree K], and g is gravitational acceleration [m s^{-2}] = 9.807 (Allen et al., 2007). Lastly, planetary boundary layer height was assumed at 2 km for all sites.

We note that not all EC sites include the necessary inputs to parameterize the Kljun et al. (2015) model, even after the above-mentioned estimation techniques. Of the full dataset evaluated, 87 sites had the necessary inputs, 69 of which are sites that were included in the benchmark ET dataset.

1.4.4. Static versus dynamic footprint comparison methods

We assessed the differences in spatial coverage between static and dynamic flux footprints. To do so, we discretized the dynamic footprints such that raster cells with a weight of 0.01 and above are considered to be within the tower footprint, which is likely a conservative limit (Chu et al., 2021), and cells with a weight below that value are outside of the footprint. Then we quantified the ability of the 3 × 3 and 7 × 7-pixel static footprints to predict the discretized dynamic footprint with a confusion matrix at each pixel. The confusion matrix analysis lets us address two important questions: (1) what fraction of the dynamic footprint is captured by the static grids on average? and (2) what fraction of the static grid is also part of the dynamic footprint? Fig. 5 is helpful to understand the confusion matrix analysis in terms of these two questions. Addressing question 1 measures the congruence between the two footprint methods, whereas question 2 gives some measure to the placement accuracy of the static footprints and whether total area tends to be smaller or larger in comparison to the dynamic footprints.

2. Results and discussion

2.1. Closure results

Average daily energy imbalance across all EC sites initially ingested was -17% during the growing season (n = 251 sites), i.e., turbulent fluxes accounted for only 83% of the available energy on average, and -22% during the non-growing season (n = 221 sites). Variations in closure and magnitudes of imbalance were much higher in the non-growing season relative to the growing season, which is expected in part because of the lower flux rates recorded in the winter that amplify the energy balance ratio (Eshonkulov et al., 2019). The energy imbalance results are similar to those found by Wilson et al. (2002) who analyzed 50 FLUXNET sites and found a 21% underestimation of turbulent fluxes on average. FLUXNET sites, however, are typically chosen from carefully sited and maintained EC sites, whereas we included all sites available to us for this analysis, some of which may not be particularly well suited for the EC technique or may have other sources of error. Another difference with our study compared to most other large scale energy imbalance studies, including Wilson et al. (2002), is that we conduct regression analysis and closure calculations at daily averaging periods, whereas half-hourly is most common in previous studies. Other studies that utilized half-hourly flux data report slightly higher energy imbalance, around 20-30% (e.g., Stoy et al., 2013; Eshonkulov et al., 2019). The reduction of closure imbalance at the daily timescale is expected and partly attributable to diurnal phase lags and hysteresis in fluxes caused by heat storage changes from day to night, which do not cancel out at half-hourly and hourly timescales (e.g., Li et al., 2008; Gao et al., 2010; Leuning et al., 2012; Dhungel et al., 2021).

When only evaluating flux sites included in the final benchmark dataset, which passed energy balance closure criteria for average growing season closure error of < 25% and non-growing season closure error of < 40%, we can make a more consistent comparison to the FLUXNET closure results. Then we find the mean energy balance closure across sites (n = 179) increases to 88% during the growing season and 86% during the non-growing season (Fig. 6).

Variation in closure based on land use / land cover is apparent in this EC dataset; for example, croplands and mixed forests exhibit the lowest relative average closure. Energy imbalance in agricultural zones can be partially explained by the fact that the conditions are often different from adjacent land areas in terms of water availability and advection, and these differences can cause losses or gains in latent and sensible heat flux that are not accounted for in the flux measurements. For example, irrigated agricultural plots that are surrounded by relatively large areas of semi-arid to arid steppe and desert in the western United States may be affected by incoming advection of hot and dry air masses, resulting in increased LE flux while decreasing H (French et al., 2012). For the sites included in the benchmark ET dataset, average growing season closure values for evergreen forest, grasslands, shrublands, and wetland/riparian sites were between 89% and 96%. The broad range in closure for grassland sites may be partially attributed to lower friction velocities (Rigden et al., 2018), which can result in insufficient convection for accurate measurement with the EC technique, particularly during stable, low-wind conditions (often at night) with poorly developed turbulence. These conditions can promote other circulations, such as gravitational drainage flows and underestimated turbulent fluxes (Aubinet et al., 2000; Barr et al., 2006). Average growing season energy balance closure values across crop subgroups including annual crops, vegetable crops, orchards, and vineyards were between 85% and 87%, with orchards having the poorest closure on average for agricultural sites.

An important question regarding energy balance closure error is whether latent and sensible heat flux measurements are biased

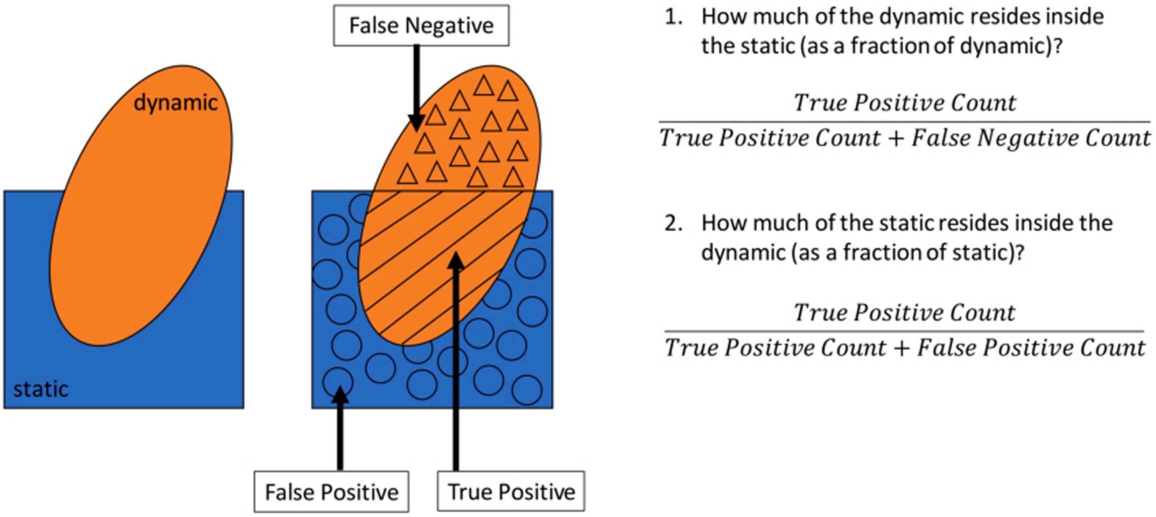
False Negative
dynamic
static
False Positive
True Positive
1. How much of the dynamic resides inside the static (as a fraction of dynamic)?
$\frac{True\ Positive\ Count}{True\ Positive\ Count + False\ Negative\ Count}$
2. How much of the static resides inside the dynamic (as a fraction of static)?
$\frac{True\ Positive\ Count}{True\ Positive\ Count + False\ Positive\ Count}$

Fig. 5. Schematic showing two approaches used to quantify the spatial relation between temporally static flux footprints based on wind rose diagrams and dynamic hourly modeled flux footprints.

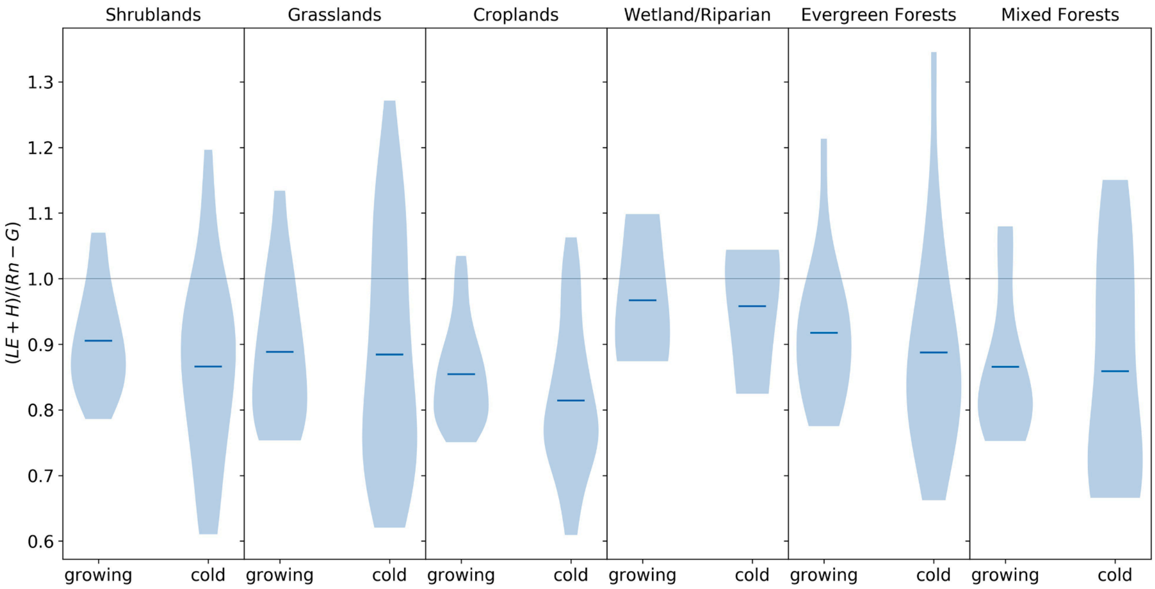

Fig. 6. Violin plot of energy balance ratios calculated from daily average energy balance variables across EC sites that passed energy balance closure criteria for the benchmark ET dataset (n=179 sites) grouped by their general land cover classifications and CGDD defined growing seasons. Horizontal dark blue bars show the average.

consistently in time and or space. The EBR and Bowen ratio methods assume consistency and enforce closure by multiplying LE and H by equal factors. Multiple linear regression (MLR) on energy balance components indicates that, assuming available energy measurements are accurate (eq. #1), LE and H are underestimated by 14% and 10% on average respectively (Table 2). When assuming R_n as the independent variable (eq. #2) results are similar (Table 3), with LE and H being underestimated by 12% and 10% respectively. Accuracy metrics for both regressions indicate that the models are robust with average r^2 values ranging from 0.79 to 0.91, and root-mean-square error (RMSE) between 15% and 26% of the average of the dependent variable (Tables 2 and 3). A dependent t-test for paired samples was calculated on regression results for LE and H coefficients. In the case where R_n – G was the dependent variable, the t-statistic was 2.3 with a p-value of 0.02, indicating insufficient evidence that the average coefficients for LE and H are identical, assuming α=0.05. However, in the regression using only R_n as the dependent variable (which is often a more reliable measurement than soil heat flux) the t-statistic was 1.5 with a p-value of 0.12, indicating (again assuming α=0.05) the mean LE and H coefficients are not significantly different. Supplementary material Fig. S11 shows the estimated probability density functions for LE and H coefficients from both regression models (eq. #1 and eq. #2) using the results from all (172) stations used in the regression analysis. Average MLR results were also differentiated by land cover type, and Tables 2 and 3 show that the difference in LE and H coefficients is highest in grasslands and wetland/riparian sites and most similar for shrubland sites. The slightly larger level of LE underestimation relative to H in grassland sites that we observe was also noted by Widmoser and Wohlfahrt (2018).

Table 2
Average results from multiple linear regression using available energy (Rn – G) as the dependent variable; results from flux sites were grouped by their general land cover type before averaging.

Land cover type	LE coef.	H coef.	r^2	RMSE [W m^{-2}]	Rn-G [W m^{-2}]	N sites
Croplands	1.16	1.12	0.83	18.7	109	68
Evergreen Forests	1.11	1.06	0.81	25.1	106	20
Grasslands	1.18	1.03	0.88	16.1	79	29
Mixed Forests	1.16	1.08	0.84	22.1	86	17
Shrublands	1.07	1.11	0.89	15.5	88	32
Wetlands	1.03	1.24	0.79	23.5	124	6

Table 3
Average results from multiple linear regression using Rn as the dependent variable; results from flux sites were grouped by their general land cover type before averaging.

Land cover type	LE coef.	H coef.	G coef.	r^2	RMSE [W m^{-2}]	Rn [W m^{-2}]	N sites
Croplands	1.16	1.13	1.04	0.86	18.0	107	68
Evergreen Forests	1.09	1.06	1.44	0.84	24.3	100	20
Grasslands	1.17	1.03	1.25	0.91	15.8	79	29
Mixed Forests	1.14	1.1	1.63	0.86	21.6	87	17
Shrublands	1.04	1.07	1.31	0.91	14.7	90	32
Wetlands	1.01	1.26	1.36	0.83	22.5	124	6

Similar average MLR results using R_n – G as compared to those using R_n as the dependent variable (Tables 2 and 3, respectively) indicate that closure errors in LE and H are not sensitive to G. This is intuitive considering G measurements are commonly less seasonally variable and of lower magnitude compared to R_n.

In addition to similar MLR coefficient values for LE and H, there is a weak positive relation (significance level, α=0.1) between the coefficients (Fig. 7). It is important for the EBR closure correction that error biases in LE and H follow each other. The biases in LE and H tend to be positively correlated in shrublands and grasslands, whereas wetland/riparian sites show a weak (not significant with α=0.1) negative relationship. Wetland and riparian sites in the dataset were often in the western United States and surrounded by more arid regions, which can create a microclimate of downward heat flux due to advection from adjacent dry and warm air masses and closing the energy balance requires a negative shift in H. Results of eq. #2, where R_n is the dependent variable, indicate that G coefficients vary much more than LE and H (Fig. S13). However, when G coefficients are near 1 (non-influencing), LE and H coefficients cluster around 1.1. A site-specific closure energy balance closure approach that also considers turbulence and advective conditions might improve results, but the overall similarity between LE and H coefficients indicates that the EBR closure correction method is reasonable.

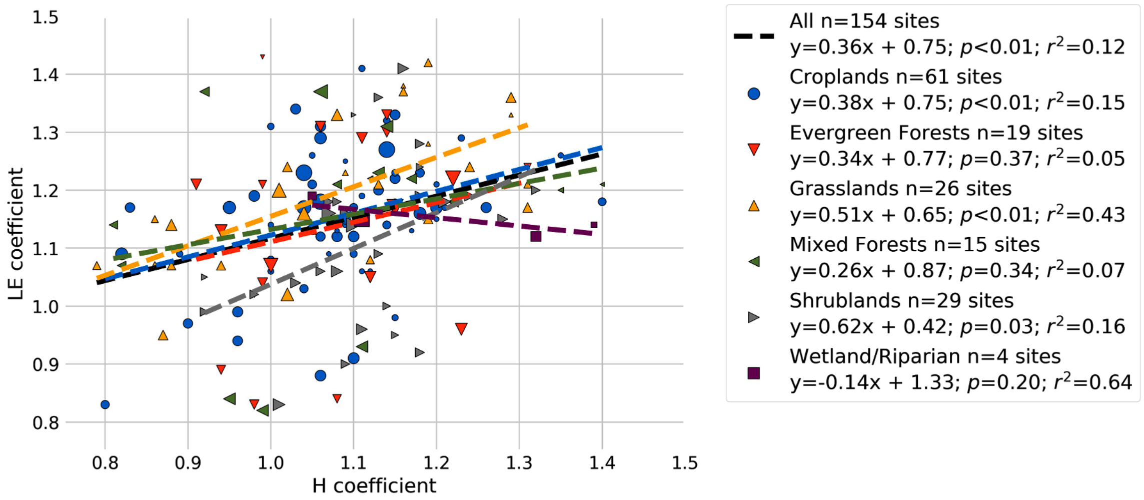

Fig. 7. Coefficients of LE versus H per EC site from multiple linear regression of surface energy balance assuming R_n – G as the dependent variable. Regression results including the slope and intercept, number of flux sites used, the f statistic p-value, and the coefficient of determination (r^2) are displayed in the legend for each land cover grouping. The sizes of the plot symbols are proportional to the sample size at each site, i.e., the number of days used in the regression. Supplementary material Fig. S12 shows the same results but with each land cover group displayed in separate plots.

2.2. Dynamic versus static footprint comparison

Overall, static grid footprints based on average daytime wind direction and speed do well at capturing the predicted flux source area as defined by the dynamic flux footprints. The larger 7 × 7 (210 m) gridded footprints included an average of 74% (daily) and 83% (monthly) of true positive pixels as a fraction of the dynamic footprint across all sites. The longer fetch distance and spatial extent of flux area indicated by these results is similar to what was found by Chu et al. (2021) who applied the Kljun et al. (2015) flux footprint model to a wide selection of EC sites across North America. The smaller 3 × 3 (90 m) static footprints capture an average of 29% (daily) and 34% (monthly) of dynamic footprint predictions (Fig.s 8 and 9). The fraction of true positive pixels to the static grids was significantly higher in the 3 × 3 grids with an average of 72% (daily) and 92% (monthly), signifying that while most of the 3 × 3 area falls within the dynamic footprint, it is not large enough to capture the total flux footprint area over time as compared to the 7 × 7 area. On the other hand, the larger 7 × 7 grids have an average fraction of true

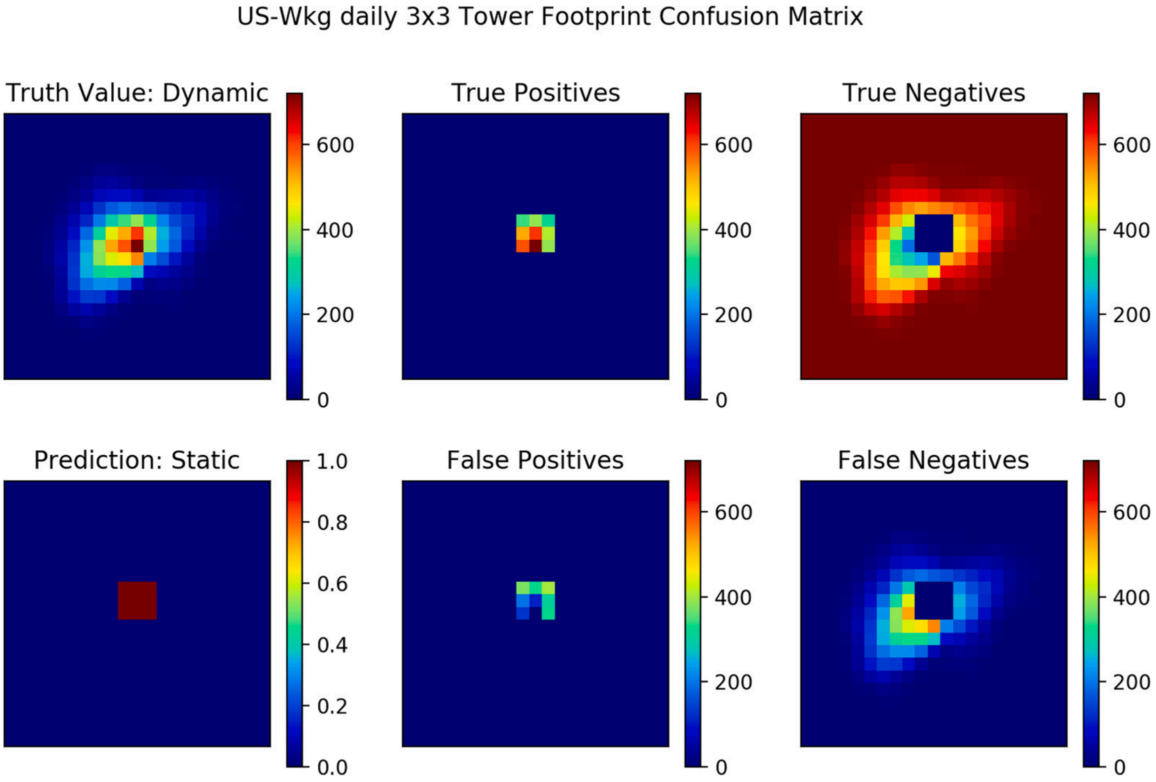

Fig. 8. Example confusion matrix footprint results for the AmeriFlux site US-Wkg (a native grassland site in southern Arizona) comparing the static 3 × 3 pixel grid with the daily dynamic footprints. Counts of pixels over all days on record are shown for all subplots except the “Prediction: Static” subplot, which only shows the location of the static 3 × 3 pixel footprint.

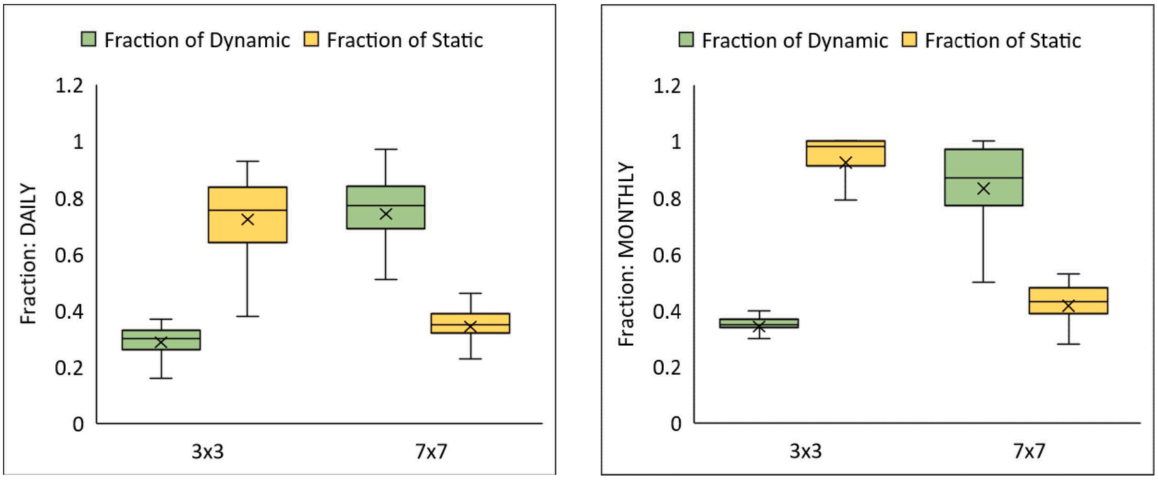

Fig. 9. Boxplots showing distribution of static-to-dynamic footprint comparison using daily (left) and monthly (right) developed from hourly ETo weighted footprints as the dynamic footprint. Fraction of dynamic (green) is the amount of the dynamic footprint that falls within a static grid as a fraction of the total dynamic footprint whereas fraction of static (yellow) is the fraction of the static that includes the dynamic. Boxplots are composed of average fractions across all flux sites, whiskers represent min/max, boxes show IQR, horizontal lines inside boxes show the median, and the x shows the mean.

positives of 34% (daily) and 42% (monthly) (Fig.s 8 and 9), meaning that on average, much of the larger grids contain pixels that are not part of the Kljun et al. (2015) footprint prediction. Because 5 × 5 grids were not generated for all stations, results are only shown for 3 × 3 and 7 × 7 grids in comparison with dynamic footprints. 5 × 5 results from a subset of sites did lie between the 3 × 3 and 7 × 7 grid results.

We found that the range of 3 × 3 true positive results between EC sites at the monthly timescale was smaller than the daily footprint result (Fig. 9). This was anticipated as daily footprint fluctuations expand beyond the 3 × 3 static footprint more often when compared to the 5 × 5 or 7 × 7 static footprints, which can better capture day-to-day variability. We also note that the interquartile range (IQR) and overall range of true positive pixels identified by the dynamic footprint that are captured by the 7 × 7 static footprints is greater than 3 × 3 footprints. This indicates differences in site-to-site footprints, and more specifically, the larger area of the 7 × 7 footprints contains more pixels and more potential to capture the temporal variability in the dynamic footprint coverage, whereas the smaller grids near the tower see less of the actual footprint and contain a more constant proportion of it over time. The range of true positive pixels as a fraction of the dynamic footprints also varied by land cover type, although whether this variability is due to land cover properties or from varying wind dynamics across sites is not clear (Fig. 9).

Overall, the footprint comparisons indicate a paradox. On one hand, the 7 × 7 grids have higher true positive counts than do the 3 × 3 (and 5 × 5) grids. However, the average fraction of the 3 × 3 grids that contain true positives was higher and over 90% at monthly timesteps, indicating that the 3 × 3 grids almost always lie within the dynamic footprint and contain very few unimportant pixels (Fig. 9). Therefore, use of the 3 × 3 grids is warranted if, for example, a larger grid may include nonrepresentative surfaces that have different reflectance properties, such as open water, buildings, roads, and paved surfaces. Because the static footprints give equal weighting to all RSET pixels within the grid, the use of the larger static footprints may not be warranted for sampling RSET pixels when nonrepresentative surfaces lie within them, particularly around the margins of the grid where the actual footprint may rarely overlap with the pixels. Ultimately, the capture of most of the important pixels by the 7 × 7 grids, as defined by the dynamic daily and monthly hourly weighted footprints, outweighs the benefit of a smaller grid size except for sites with heterogeneous surroundings where it would be best to utilize smaller 5 × 5 or 3 × 3 grids. The measurement and canopy height should also be used to guide decisions to revert to smaller sized grids. For example, typical fetch distances for Bowen ratio instrumented sites are likely shorter than EC systems (Stannard, 1997), and even 3 × 3 grids may be large for RSET sampling at lysimeter sites (Kustas et al., 2015). In locations with consistently long fetches, use of smaller grids may be inappropriate and a manual adjustment of the larger footprint should be applied to avoid nonrepresentative pixels at heterogeneous sites.

The effect of footprint choice on the accuracy of sampled RSET estimates is beyond the scope of this paper. However, in the context of tradeoffs between footprint size and representativity, it should be noted that while the larger 7 × 7 static grids are more likely to contain most of the actual footprint, they have a higher potential of sampling pixels that are not part of the actual footprint over any given period. If the surface heterogeneity is high around the tower, which is not uncommon within AmeriFlux sites (Chu et al., 2021), then probability increases of sampling nonrepresentative pixels that may cause different sized grids to yield very different RSET ET estimates (Fisher et al., 2020). If the footprint includes many, often non-contributing pixels that are of a different land use / land cover, they should be avoided by slightly shifting the footprint to avoid these pixels or using a smaller domain.

2.3. Limitations and future directions

The OpenET benchmark ET dataset consists of 194 EC sites that are well distributed across the CONUS. However, a major objective moving forward is to evolve and expand the dataset by increasing its coverage in time and space. Towards this end, a data pipeline has been developed to ingest additional EC datasets, including those from AmeriFlux or from site PIs, as they become available. Having as much temporal and spatial coverage as possible is important to facilitate robust accuracy assessment of RSET models across the full range of agricultural crops and land cover and climate conditions across the United States. Some specific regions of interest for additional ET stations that currently are not well represented in the dataset include agricultural regions of the Pacific Northwest and the Upper Colorado River Basin.

Currently, 2-dimensional dynamic flux footprints have been developed for 87 EC sites. Dynamic footprints are planned to be developed for the remaining sites as additional input data become available. Other adjustments and improvements may be applied to flux footprints (e.g., as a result of a sensitivity analysis of RSET pixel sampling on footprint methods).

Many EC sites that were included in the benchmark dataset did not pass energy balance closure requirements or other QA/QC checks. These sites were not included in the benchmark dataset for the initial OpenET intercomparison and accuracy assessment; however, these sites can be reassessed in future analyses using recently collected data. In the future, new methods for energy balance closure correction, gap-filling, data filtering, and other corrections could be applied to energy balance components to improve energy balance closure and reduce uncertainty in the benchmark ET dataset. For example, it is not always clear whether measurements of soil heat fluxes at EC sites that are uploaded to networks like AmeriFlux have accounted for soil heat storage above the flux plate. At sites where heat storage has not been accounted for in the measurement of soil heat flux, investigating and potentially estimating soil heat storage as a function of soil temperature, moisture content, and soil properties would be useful (Purdy et al., 2016).

Although it is beyond the scope of this paper, we plan to further explore energy imbalance at the EC sites at daily and sub-daily scales; for example, by comparing closure based on time of day or using daytime averaging periods for energy balance assessment. It would be insightful to contrast energy balance error with micrometeorological, advective, and stability conditions, similar to the analysis of Bambach et al. (2022), and connect it to regional differences such as land use and climate. Identification of factors and conditions that are related to EC energy balance in measurements can be used to help improve data quality assessments and inform methods to adjust turbulent fluxes to compensate for energy balance closure error.

The OpenET benchmark dataset and tools were developed for RSET model evaluation; however, the wide spatio-temporal coverage and standardized methodologies of the dataset facilitates many potential applications and multidisciplinary studies. For example, the daily and monthly ET and/or meteorological data can be used for input, calibration, and/or evaluation of regional or large scale hydrologic or land surface models (e.g., Swenson et al., 2019) that are often validated against streamflow measurements alone.

3. Summary

The OpenET benchmark dataset is primarily a collection of post-processed daily and monthly ET data that have been corrected for energy imbalance along with open-source tools for data provenance of 194 ET stations (179 of which are eddy covariance systems) across the CONUS. Data were combined from multiple providers including AmeriFlux, the USDA, USGS, and university partners. All EC data underwent the same gap-filling, time aggregation, energy balance closure corrections, and visual inspection data quality checks and data filtering.

Energy balance closure analysis was conducted on a large set of EC sites (251), and we found average levels of energy balance closure error to be near the lower end of the typical range reported for EC sites. This is primarily caused by using daily average fluxes as opposed to half-hourly. Multiple linear regression of daily averaged fluxes and energy provided evidence that sensible and latent heat fluxes from most EC sites tend to be underestimated by similar magnitudes on average. This supported our decision to apply the EBR closure technique at daily timescales. We also found variability in closure based on land cover type and identified that conducting further research on other atmospheric and physical factors that may control, or be related to, the energy imbalance at flux towers would be useful.

Evaluation of flux footprints used for sampling model ET pixels at each ground station included static grids (e.g., 3 × 3 or 30-meter pixel grids) based on the daytime long-term wind speed and direction and, data permitting, daily and monthly dynamic flux footprints composed from hourly footprints and weighted by hourly reference ET. An in-depth comparison of these methods at 87 stations revealed that while larger (7 × 7 or 210 m) grids captured most of the dynamic footprints, the smaller static footprints (3 × 3 or 90 m) had a higher fraction of their area fall within the corresponding dynamic flux footprint extent. Both results signified accurate placement of gridded footprints with respect to the physically based footprint model.

A subset of 194 of more than 300 initial ET stations were selected based on data coverage, energy balance closure criteria, and other data quality checks to comprise this benchmark ET dataset. The data and methods described here are being used to conduct a large-scale intercomparison and accuracy assessment of the six satellite-based models that have been implemented within OpenET. The ET dataset and associated tools have other potential uses for a variety of atmospheric and hydrologic scientific investigations. Collaboration by the community to improve the dataset and the associated open-source tools would be beneficial.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Post-processed ET data are available in the accompanying "Data in Brief" article. Code is open-source on GitHub. Other data are in the supplementary material or are available on request.

Acknowledgements

Development of OpenET and this work is supported by the S.D. Bechtel, Jr. Foundation; the Gordon and Betty Moore Foundation; the Walton Family Foundation; Lyda Hill Philanthropies; the Windward Fund; the Keith Campbell Foundation for the Environment; the North, Central, and South Delta Water Agencies; the NASA Applied Science Program and the NASA Western Water Applications Office; the U.S. Geological Survey (USGS) Landsat Science Team; the California State University Agricultural Research Institute; the Idaho Agricultural Experiment Station and Nebraska Agricultural Experiment Station. In-kind support is provided by partners in the agricultural and water management communities, Google Earth Engine, and the Water Funder Initiative. In addition, funding for AmeriFlux data resources was provided by the U.S. Department of Energy's Office of Science. This research includes contributions from the Long-Term Agroecosystem Research (LTAR) network. LTAR is supported by the United States Department of Agriculture (USDA). We acknowledge and thank the long-term data collection efforts by the AmeriFlux program, the USDA Agricultural Research Service, the USGS Nevada Water Science Center, and the Delta-Flux network. We acknowledge and thank the following AmeriFlux sites and team members for their data records: US-A32, US-A74, US-ADR, US-AR1, US-ARb, US-ARc, US-ARM, US-Aud, US-Bi1, US-Bi2, US-Bkg, US-Blk, US-Blo, US-Bo1, US-Br1, US-Br3, US-Ced, US-CMW, US-CRT, US-Ctn, US-CZ3, US-Dix, US-Dk1, US-Dk2, US-Esm, US-Fmf, US-FPe, US-FR2, US-Fuf, US-Fwf, US-GLE, US-GMF, US-Goo, US-Hn2, US-Hn3, US-IB1, US-IB2, US-Jo2, US-KLS, US-KM4, US-KS2, US-LS1, US-MC1, US-Me1, US-Me2, US-Me5, US-Me6, US-Mj1, US-Mj2, US-MMS, US-MOz, US-NC2, US-NC3, US-NC4, US-Ne1, US-Ne2, US-Ne3, US-NR1, US-Oho, US-Ro1, US-Ro2, US-Ro3, US-Ro4, US-Ro5, US-Ro6, US-Rwe, US-Rwf, US-Rws, US-SCg, US-SCs, US-SCw, US-SdH, US-Skr, US-Slt, US-Sne, US-SO2, US-SO3, US-SO4, US-SP2, US-SP3, US-SP4, US-SRC, US-SRG, US-SRM, US-Srr, US-SRS, US-Tw2, US-Tw3, US-Twt, US-Var, US-WBW, US-WCr, US-Wkg, US-xAE, US-xDC, US-xDL, US-xDS, US-xJR, US-xNG, US-xNW, US-xRM, US-xSB, US-xSL, US-xST, US-xUN, US-xYE, US-AR2, US-ChR, US-CZ2, US-DFC, US-FR3, US-GBT, US-HB3, US-KUT, US-LS2, US-LWW, US-MtB, US-ONA, US-Pon, US-Rls, US-Rms, US-SFP, US-Snd, US-Snf, US-Syv, US-UiA, US-Whs, US-xTR.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.agrformet.2023.109307.

References

Abatzoglou, J.T., 2013. Development of gridded surface meteorological data for ecological applications and modelling. Int. J. Climatol. 33 (1), 121–131.

Allen, R.G., Robison, C.W., 2007. Evapotranspiration and Net Irrigation Water Requirements for Idaho. University of Idaho reported submitted to Idaho Department of Water Resources. Retrieved from. www.kimberly.uidaho.edu/ETIdaho.

Allen, R.G., Pereira, L.S., Howell, T.A., Jensen, M.E., 2011. Evapotranspiration information reporting: I. Factors governing measurement accuracy. Agric. Water Manage. 98 (6), 899–920.

Allen, R.G., Tasumi, M., Trezza, R., 2007. Satellite-based energy balance for mapping evapotranspiration with internalized calibration (METRIC) – model. ASCE *J. Irrigat. Drainage Eng.* 133 (4), 380–394.

Allen, R.G., Walter, I.A., Elliott, R.L., Howell, T.A., Itenfisu, D., Jensen, M.E., Snyder, R. L., 2005. The ASCE Standardized Reference Evapotranspiration Equation. American Society of Civil Engineers. https://doi.org/10.1061/9780784408056.

Arya, S.P., 1998. Introduction to Micrometeorology. Academic Press, San Diego.

Aubinet, M., Grelle, A., Ibrom, A., Rannik, Ü., Moncrieff, J., Foken, T., Kowalski, A.S., Martin, P.H., Berbigier, P., Bernhofer, C., Clement, R., Elbers, J., Granier, A., Grünwald, T., Morgenstern, K., Pilegaard, K., Rebmann, C., Snijders, W., Valentini, R., Vesala, T., 2000. Estimates of the annual net carbon and water exchange of forest: the EUROFLUX methodology. Adv. Ecol. Res. 30, 113–175.

Bayat, B., Camacho, F., Nickeson, J., Cosh, M., Bolten, J., Vereecken, H., Montzka, C., 2021. Toward operational validation systems for global satellite-based terrestrial essential climate variables. Int. J. Appl. Earth Obs. Geoinf. 95, 102240 https://doi.org/10.1016/J.JAG.2020.102240.

Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., Anthoni, P., Bernhofer, C., Davis, K., Evans, R., Fuentes, J., 2001. FLUXNET: A new tool to study the temporal and spatial variability of ecosystem-scale carbon dioxide, water vapor, and energy flux densities. Bull. Am. Meteorol. Soc. 82 (11), 2415–2434.

Baldocchi, D., 2014. Measuring fluxes of trace gases and energy between ecosystems and the atmosphere–the state and future of the eddy covariance method. Global Change Biol. 20 (12), 3600–3609.

Bambach, N., Kustas, W., Alfieri, J., Prueger, J., Hipps, L., McKee, L., Castro, S.J., Volk, J., Alsina, M.M., McElrone, A.J., 2022. Evapotranspiration uncertainty at micrometeorological scales: the impact of the eddy covariance energy imbalance and correction methods. Irrigat. Sci. 1432, 1319. https://doi.org/10.1007/s00271-022-00783-1.

Barr, A.G., Morgenstern, K., Black, T.A., McCaughey, J.H., Nesic, Z., 2006. Surface energy balance closure by the eddy-covariance method above three boreal forest stands and implications for the measurement of the CO2 flux. Agric. For. Meteorol. 140, 322–337.

Chu, H., Baldocchi, D.D., Poindexter, C., Abraha, M., Desai, A.R., Bohrer, G., Arain, M.A., Griffis, T., Blanken, P.D., O'Halloran, T.L., Thomas, R.Q., 2018. Temporal dynamics of aerodynamic canopy height derived from eddy covariance momentum flux data across North American flux networks. Geophys. Res. Lett. 45 (17), 9275–9287.

Chu, H., Luo, X., Ouyang, Z., Chan, W.S., Dengel, S., Biraud, S.C., Torn, M.S., Metzger, S., Kumar, J., Arain, M.A., Arkebauer, T.J., 2021. Representativeness of Eddy-Covariance flux footprints for areas surrounding AmeriFlux sites. Agric. For. Meteorol. 301, 108350.

Dhungel, R., Aiken, R., Evett, S.R., Colaizzi, P.D., Marek, G., Moorhead, J.E., Baumhardt, R.L., Brauer, D., Kutikoff, S., Lin, X, 2021. Energy imbalance and evapotranspiration hysteresis under an advective environment: evidence from lysimeter, eddy covariance, and energy balance modeling. Geophys. Res. Lett. 48, e2020GL091203 https://doi.org/10.1029/2020GL091203.

Eshonkulov, R., Poyda, A., Ingwersen, J., Wizemann, H.D., Weber, T.K., Kremer, P., Högy, P., Pulatov, A., Streck, T., 2019. Evaluating multi-year, multi-site data on the energy balance closure of eddy-covariance flux measurements at cropland sites in southwestern Germany. Biogeosciences 16 (2), 521–540.

Evett, S., Howell, T., Schneider, A.D., Copeland, K.S., Dusek, D.A., Brauer, D., Tolk, J., Marek, G., Marek, T., Gowda, P., 2016. The Bushland weighing lysimeters: a quarter century of crop ET investigations to advance sustainable irrigation. Trans. ASABE 59, 163–179.

Finnigan, J.J., Clement, R., Malhi, Y., Leuning, R., Cleugh, H.A., 2003. A re-evaluation of long-term flux measurement techniques part I: averaging and coordinate rotation. Boundary Layer Meteorol. 107 (1), 1–48.

Fisher, J.B., Baldocchi, D.D., Misson, L., Dawson, T., Goldstein, A.H., 2007. What the towers don't see at night: Nocturnal sap flow in trees and shrubs at two AmeriFlux sites in California. Tree Physiol. 27 (4), 597–610.

Fisher, J.B., Tu, K.P., Baldocchi, D.D., 2008. Global estimates of the land–atmosphere water flux based on monthly AVHRR and ISLSCP-II data, validated at 16 FLUXNET sites. Remote Sens. Environ. 112 (3), 901–919.

Fisher, J.B., Melton, F., Middleton, E., Hain, C., Anderson, M., Allen, R., McCabe, M.F., Hook, Baldocchi, D., Townsend, P.A., Kilic, A., Tu, K., Miralles, D.D., Perret, J., Lagouarde, J.-P., Waliser, D., Purdy, A.J., French, A., Schimel, D., Famiglietti, J.S., Stephens, G., Wood, E.F., 2017. The future of evapotranspiration: global requirements for ecosystem functioning, carbon and climate feedbacks, agricultural management, and water resources. Water Resour. Res. 53 (4), 2618–2626.

Fisher, J.B., Lee, B., Purdy, A.J., Halverson, G.H., Dohlen, M.B., Cawse-Nicholson, K., Wang, A., Anderson, R.G., Aragon, B., Arain, M.A., Baldocchi, D.D., Baker, J.M., Barral, H., Bernacchi, C.J., Bernhofer, C., Biraud, S.C., Bohrer, G., Brunsell, N., Cappelaere, B., Castro-Contreras, S., Chun, J., Conrad, B.J., Cremonese, E., Demarty, J., Desai, A.R., De Ligne, A., Foltýnová, L., Goulden, M.L., Griffis, T.J., Grünwald, T., Johnson, M.S., Kang, M., Kelbe, D., Kowalska, N., Lim, J.H., Maïnassara, I., McCabe, M.F., Missik, J.E.C., Mohanty, B.P., Moore, C.E., Morillas, L., Morrison, R., Munger, J.W., Posse, G., Richardson, A.D., Russell, E.S., Ryu, Y., Sanchez-Azofeifa, A., Schmidt, M., Schwartz, E., Sharp, I., Šigut, L., Tang, Y., Hulley, G., Anderson, M., Hain, C., French, A., Wood, E., Hook, S., 2020. ECOSTRESS: NASA's next generation mission to measure evapotranspiration from the International Space Station. Water Resour. Res. 56 (4), 1–20.

Foken, T., 2008. The energy balance closure problem: an overview. Ecol. Appl. 18 (6), 1351–1367. https://doi.org/10.1890/06-0922.1.

French, A.N., Alfieri, J.G., Kustas, W.P., Prueger, J.H., Hipps, L.E., Chávez, J.L., Evett, S. R., Howell, T.A., Gowda, P.H., Hunsaker, D.J., Thorp, K.R., 2012. Estimation of surface energy fluxes using surface renewal and flux variance techniques over an advective irrigated agricultural site. Adv. Water Res. 50, 91–105. https://doi.org/10.1016/j.advwatres.2012.07.007.

Gao, Z., Horton, R., Liu, H.P., 2010. Impact of wave phase difference between soil surface heat flux and soil surface temperature on soil surface energy balance closure. J. Geophys. Res. 115 (D16).

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google earth engine: planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27.

Harrison, L.P., 1963. Fundamental concepts and definitions relating to humidity. In: Wexler, A. (Ed.), Humidity and Moisture. Reinhold, New York. Vol. 3.

Högström, U., Smedman, A.S., 2004. Accuracy of sonic anemometers: laminar wind-tunnel calibrations compared to atmospheric in situ calibrations against a reference instrument. Boundary Layer Meteorol. 111 (1), 33–54.

Huntington, J.L., Allen, R., 2009. Evapotranspiration and Net Irrigation Water Requirements for Nevada. Nevada State Engineer's Office Publication, p. 266.

Jensen, M.E., Allen, R.G, 2016. Direct Penman-Monteith and aerodynamic energy balance equations. In: Jensen, M.E., Allen, R.G. (Eds.), Evaporation, Evapotranspiration, and Irrigation Water Requirements. https://doi.org/10.1061/9780784414057.

Kaimal, J., Finnigan, J., 1994. Atmospheric Boundary Layer Flows: Their Structure and Measurement. Oxford University Press, Oxford, p. 289.

Kljun, N., Calanca, P., Rotach, M.W., Schmid, H.P., 2015. A simple two-dimensional parameterisation for Flux Footprint Prediction (FFP). Geosci. Model Develop. 8 (11), 3695.

Kormann, R., Meixner, F.X., 2001. An analytical footprint model for non-neutral stratification. Boundary Layer Meteorol. 99 (2), 207–224.

Kristensen, L., Mann, J., Oncley, S.P., Wyngaard, J.C., 1997. How close is close enough when measuring scalar fluxes with displaced sensors? J. Atmos. Oceanic Technol. 14 (4), 814–821.

Kustas, W.P., Anderson, M.C., Alfieri, J.G., Knipper, K., Torres-Rua, A., Parry, C.K., Nieto, H., Agam, N., White, W.A., Gao, F., McKee, L., 2018. The grape remote sensing atmospheric profile and evapotranspiration experiment. Bull. Am. Meteorol. Soc. 99 (9), 1791–1812.

Kustas, W.P., Alfieri, J.G., Evett, S., Agam, N., 2015. Quantifying variability in field-scale evapotranspiration measurements in an irrigated agricultural region under advection. Irrigat. Sci. 33 (5), 325–338.

Leuning, R., Van Gorsel, E., Massman, W.J., Isaac, P.R., 2012. Reflections on the surface energy imbalance problem. Agric. For. Meteorol. 156, 65–74.

Li, Q., Zhang, X.Z., Shi, P.L., He, Y.T., Xu, L.L., Sun, W., 2008. Study on the energy balance closure of Alpine meadow on Tibetan Plateau. J. Natural Res. 23, 391–399.

Lin, C., Gentine, P., Frankenberg, C., Zhou, S., Kennedy, D., Li, X., 2019. Evaluation and mechanism exploration of the diurnal hysteresis of ecosystem fluxes. Agric. For. Meteorol. 278, 107642 https://doi.org/10.1016/J.AGRFORMET.2019.107642.

Massman, W.J., Lee, X., 2002. Eddy covariance flux corrections and uncertainties in long-term studies of carbon and energy exchanges. Agric. For. Meteorol. 113 (1-4), 121–144.

Mauder, M., Cuntz, M., Drüe, C., Graf, A., Rebmann, C., Schmid, H.P., Schmidt, M., Steinbrecher, R., 2013. A strategy for quality and uncertainty assessment of long-term eddy-covariance measurements. Agric. For. Meteorol. 169, 122–135. https://doi.org/10.1016/J.AGRFORMET.2012.09.006.

Melton, F.S., Huntington, J., Grimm, R., Herring, J., Hall, M., Rollison, D., Erickson, T., Allen, R., Anderson, M., Fisher, J.B., Kilic, A., Senay, G.B., Volk, J., Hain, C., Johnson, L., Ruhoff, A., Blankenau, P., Bromley, M., Carrara, W., Daudert, B., Doherty, C., Dunkerly, C., Friedrichs, M., Guzman, A., Halverson, G., Hansen, J., Harding, J., Kang, Y., Ketchum, D., Minor, B., Morton, C., Ortega-Salazar, S., Ott, T., Ozdogan, M., ReVelle, P.M., Schull, M., Wang, C., Yang, Y, 2021. OpenET: filling a critical data gap in water management for the western United States. JAWRA J. Ame. Water Res. Ass. https://doi.org/10.1111/1752-1688.12956.

Moncrieff, J., Clement, R., Finnigan, J., Meyers, T., 2004. Averaging, detrending, and filtering of eddy covariance time series. In: Lee, X., Massman, W., Law, B. (Eds.), Handbook of Micrometeorology. *Atmospheric and Oceanographic Sciences Library*. Springer, Dordrecht. https://doi.org/10.1007/1-4020-2265-4_2 vol 29.

Moore, C.J., 1986. Frequency response corrections for eddy correlation systems. Boundary Layer Meteorol. 37 (1), 17–35.

Pahlow, M., Parlange, M.B., Porté-Agel, F., 2001. On Monin–Obukhov similarity in the stable atmospheric boundary layer. Boundary Layer Meteorol. 99, 225–248. https://doi.org/10.1023/A:1018909000098.

Papale, D., Reichstein, M., Aubinet, M., Canfora, E., Bernhofer, C., Kutsch, W., Longdoz, B., Rambal, S., Valentini, R., Vesala, T., Yakir, D., 2006. Towards a standardized processing of Net Ecosystem Exchange measured with eddy covariance technique: algorithms and uncertainty estimation. Biogeosciences 3 (4), 571–583.

Pastorello, G., Trotta, C., Canfora, E., et al., 2020. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. Scientific Data 7 (225), 1–27. https://doi.org/10.1038/s41597-020-0534-3.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: Machine learning in Python. J. Machine Learn. Res. 12 (85), 2825–2830.

Pennypacker, S., Baldocchi, D., 2016. Seeing the fields and forests: application of surface-layer theory and flux-tower data to calculating vegetation canopy height. Boundary Layer Meteorol. 158, 165–182. https://doi.org/10.1007/s10546-015-0090-0.

Prueger, J.H., Kustas, W.P., 2005. Aerodynamic methods for estimating turbulent fluxes, in micrometeorology in agricultural systems. In: Hatfield, J.L., Baker, J.L. (Eds.), Agronomy Monograph No. 47. ASA-CSSA-SSSA, Madison, WI, pp. 407–436.

Purdy, A., Fisher, J.B., Goulden, M.L., Famiglietti, J.S., 2016. Ground heat flux: an analytical review of 6 models evaluated at 88 sites and globally. J. Geophys. Res. 121 https://doi.org/10.1002/2016JG003591.

Rigden, A., Li, D., Salvucci, G., 2018. Dependence of thermal roughness length on friction velocity across land cover types: a synthesis analysis using AmeriFlux data. Agric. For. Meteorol. 249, 512–519. https://doi.org/10.1016/J.AGRFORMET.2017.06.003.

Rosenberg, N.J., Blad, B.L., Verma, S.B., 1983. Microclimate: the Biological Environment, 2nd Ed. John Wiley & Sons, p. 495.

Runkle, B.R.K., Rigby, J.R., Reba, M.L., Anapalli, S.S., Bhattacharjee, J., Krauss, K.W., Liang, L., Locke, M.A., Novick, K.A., Sui, R., Suvočarev, K., White, P.M., 2017. Delta-Flux: an eddy covariance network for a climate-smart Lower Mississippi Basin. Agricult. Environ. Lett. 2 (1) https://doi.org/10.2134/AEL2017.01.0003 ael2017.01.0003.

Sammis, T., Mapel, C., Lugg, D.G., Lansford, R.R., McGuckin, J.T., 1985. Evapotranspiration crop coefficients predicted using growing degree days. Trans. Am. Soc. Agricult. Eng. 28 (3), 773–780.

Smedman, A.S., 1988. Observations of a multi-level turbulence structure in a very stable atmospheric boundary layer. Boundary Layer Meteorol. 44, 231–253. https://doi.org/10.1007/BF00116064.

Stannard, D.I., 1997. A theoretically based determination of Bowen-ratio fetch requirements. Boundary Layer Meteorol. 83 (3), 375–406.

Stoy, P.C., Mauder, M., Foken, T., Marcolla, B., Boegh, E., Ibrom, A., Arain, M.A., Arneth, A., Aurela, M., Bernhofer, C., Cescatti, A., 2013. A data-driven analysis of energy balance closure across FLUXNET research sites: the role of landscape scale heterogeneity. Agric. For. Meteorol. 171, 137–152.

Stull, R.B., 1988. An Introduction to Boundary-Layer Meteorology. Kluwer Academic, Dordrecht, The Netherlands.

Swenson, S.C., Clark, M., Fan, Y., Lawrence, D.M., Perket, J., 2019. Representing intra-hillslope lateral subsurface flow in the community land model. J. Adv. Model. Earth Syst. 11, 4044–4065. https://doi.org/10.1029/2019MS001833.

Twine, T., Kustas, W.P., Norman, J., Cook, D., Houser, P., Teyers, T.P., Prueger, J.H., Starks, P., Wesely, M., 2000. Correcting Eddy-covariance flux underestimates over a grassland. Agric. For. Meteorol. 103 (3), 279–300.

Vickers, D., Mahrt, L., 2007. Observations of the cross-wind velocity variance in the stable boundary layer. Environ. Fluid Mech. 7 (1), 55–71.

Volk, J.M., Huntington, J.L., Allen, R., Melton, F., Anderson, M., Kilic, A., 2021. flux-data-qaqc: a python package for energy balance closure and post-processing of Eddy flux data. J. Open Source Software 6 (66), 3418. https://doi.org/10.21105/joss.03418.

Volk, J., Huntington, J., Melton, F., Minor, B., Wang, T., Anapalli, S., Anderson, R., Evett, S., French, A., Jasoni, R., Bambach, N., Kustas, W., Alfieri, J., Prueger, J., Hipps, L., McKee, L., Castro, S.J., Alsina, M.M., McElrone, A.J., Runkle, B., Saber, M., Sanchez, C., Tajfar, E., Anderson, M, 2022. Post-processed daily and monthly data for a benchmark CONUS eddy flux ET dataset. Data in Brief submitted.

Wilson, K., Goldstein, A., Falge, E., Aubinet, M., Baldocchi, D., Berbigier, P., Bernhofer, C., Ceulemans, R., Dolman, H., Field, C., Grelle, A., 2002. Energy balance closure at FLUXNET sites. Agric. For. Meteorol. 113 (1-4), 223–243.

Wright, J.L., 2001. Growing degree day functions for use with evapotranspiration crop coefficients. CD-ROM. American Society of Agronomy. Agronomy Abstracts.

Xia, Y., Mitchell, K., Ek, M., Cosgrove, B., Sheffield, J., Luo, L., Alonge, C., Wei, H., Meng, J., Livneh, B., Duan, Q., 2012. Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. 1. Intercomparison and application of model products. J. Geophys. Res. 117 (D3).