nature water

Analysis

https://doi.org/10.1038/s44221-023-00181-7

Assessing the accuracy of OpenET satellite-based evapotranspiration data to support water resource and land management applications

Received: 21 June 2023

Accepted: 30 November 2023

Published online: 15 January 2024

Check for updates

John M. Volk[1]✉, Justin L. Huntington[1], Forrest S. Melton[2,3], Richard Allen[4], Martha Anderson[5], Joshua B. Fisher[6], Ayse Kilic[7], Anderson Ruhoff[8], Gabriel B. Senay[9], Blake Minor[1], Charles Morton[1], Thomas Ott[1], Lee Johnson[2,3], Bruno Comini de Andrade[8], Will Carrara[2,3], Conor T. Doherty[2], Christian Dunkerly[1], MacKenzie Friedrichs[10], Alberto Guzman[2,3], Christopher Hain[11], Gregory Halverson[12], Yanghui Kang[13], Kyle Knipper[14], Leonardo Laipelt[8], Samuel Ortega-Salazar[7], Christopher Pearson[1], Gabriel E. L. Parrish[15], Adam Purdy[2,3], Peter ReVelle[7], Tianxin Wang[13] & Yun Yang[16]

Remotely sensed evapotranspiration (ET) data offer strong potential to support data-driven approaches for sustainable water management. However, practitioners require robust and rigorous accuracy assessments of such data. The OpenET system, which includes an ensemble of six remote sensing models, was developed to increase access to field-scale (30 m) ET data for the contiguous United States. Here we compare OpenET outputs against data from 152 in situ stations, primarily eddy covariance flux towers, deployed across the contiguous United States. Mean absolute error at cropland sites for the OpenET ensemble value is 15.8 mm per month (17% of mean observed ET), mean bias error is −5.3 mm per month (6%) and r^2 is 0.9. Results for shrublands and forested sites show higher inter-model variability and lower accuracy relative to croplands. High accuracy and multi-model convergence across croplands demonstrate the utility of a model ensemble approach, and enhance confidence among ET data practitioners, including the agricultural water resource management community.

Accurate evapotranspiration (ET) data are essential for assessing the surface energy and water balance, the carbon cycle and the management of water resources[1]. ET is the sum of the flux of water vapour from soil (evaporation) and through vegetation (transpiration) to the atmosphere. ET constitutes the second largest component of the terrestrial water balance, after precipitation. The usefulness of spatially contiguous mapping of ET, particularly over irrigated agricultural lands, has been amplified by drought, climate change, and high rates of human water withdrawal and agricultural consumption, leaving many aquifers and water reservoirs in the western United States at all-time-low levels[2–4]. Satellite-based remote sensing of ET (RSET) offers a powerful approach for mapping ET over large geographic regions at semi-continuous timescales[1,5,6]. Until recently, the availability of RSET data at spatial scales relevant for water resources management has been limited by cost and computational requirements.

OpenET[5] employs six state-of-the-art satellite based RSET models, that is, ALEXI/DisALEXI[7], eeMETRIC[8], geeSEBAL[9], PT-JPL[10], SIMS[11,12] and SSEBop[13], that have been widely applied and evaluated in the United

A full list of affiliations appears at the end of the paper. ✉e-mail: john.volk@dri.edu

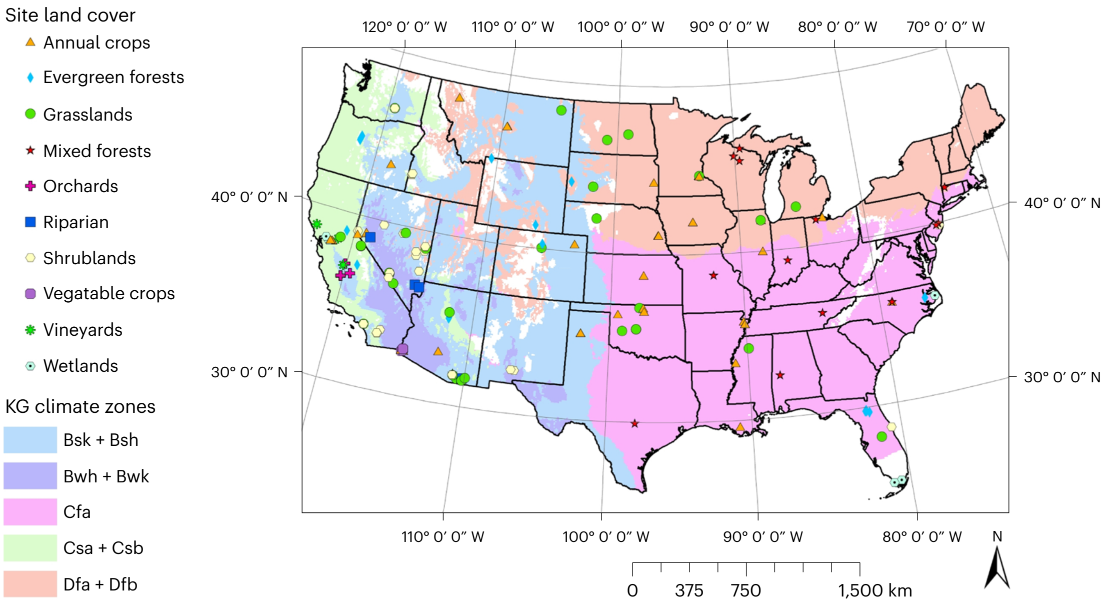


Fig. 1 | Map of in situ ET measurement sites. Map of the locations of in situ ET stations used to evaluate OpenET, including their general land cover type and Köppen–Geiger (KG) climate zones[34]. White areas represent climate zones that did not contain any cropland sites and were excluded from the analysis. Climate zone abbreviations are defined as follows: cold and hot semi-arid steppe (Bsk + Bsh); hot and cold desert (Bwh + Bwk); humid subtropical (Cfa); hot- and warm-summer Mediterranean (Csa + Csb); and hot- and warm-summer humid continental (Dfa + Dfb).

States for a range of water management and agricultural applications. The models are applied on the Google Earth Engine cloud-based platform[14] to provide historical and near real-time ET data at subfield scales (30-m pixels) over the western United States[5]. Five of the RSET models constrain components of the surface energy balance (SEB) using land surface temperature (LST) primarily derived from Landsat Collection 2, along with gridded weather data, and land cover datasets. The sixth model, SIMS, assumes well-watered conditions and computes crop coefficients based on vegetation density, derived from satellite surface reflectance values, along with a gridded soil water balance model. The models composing OpenET have been used by water managers, farmers and governmental organizations for irrigation scheduling, water accounting and allocation, and water rights administration[15–17]. The OpenET platform provides an unprecedented level of accessibility to RSET data through its public online data explorer interface—including querying satellite ET within individually vectorized field boundaries. All six RSET models in OpenET operate automatically, including any required calibrations, which permits rapid calculations for the more than 100,000 Landsat images processed so far across the 23 western-most states in the contiguous United States. As the number of applications of RSET data for sustainable land and water resources management grow, it is important for practitioners to have information on the accuracy of RSET data across land cover types, climatic zones and agricultural production practices[18].

In this Analysis, we present a large-scale benchmark assessment of the accuracy of OpenET data using a well-curated publicly archived dataset of in situ ET measurements from 152 stations (141 eddy covariance (EC) systems, 7 Bowen ratio systems and 4 lysimeters), over a variety of regions, climates and land cover types[19,20], collectively comprising ~45 years of paired model–measurement ET data (Fig. 1). The EC technique is generally viewed as the best available method for continuous measurement of in situ energy and heat flux at spatial scales that approach satellite-based retrievals[21,22], although we acknowledge the associated data uncertainties and made efforts to reduce them[19]. In addition to evaluation of individual model accuracies, we evaluated the OpenET ensemble ET value, computed as the mean of all models after flagging and removal of up to two outliers using the median absolute deviation (MAD) approach[23,24]. The generation of an ensemble value is a widely used technique to combine outputs from diverse models, each having their own behaviour[5] and random error[25–27]. It also facilitates applications such as irrigation scheduling and water rights administration, where practitioners require a single value for use in management of water resources[5]. The publicly archived in situ flux dataset allows for reproducibility and benchmarking of future OpenET model versions or other RSET data.

ET data computed from micrometeorological measurements at EC sites were obtained from a variety of sources, primarily AmeriFlux[28]. Supplementary Table 1 provides a full list of stations used in the study including land cover type, site principal investigators, Digital Object Identifiers (DOIs) and other metadata. Flux data were carefully post-processed, including gap-filling, screening for energy balance closure error and data completeness, and visual data quality assessments. Flux data that passed quality control and showed limited energy balance closure error were included in the study and underwent closure correction following the FLUXNET2015/ONEFlux approach for daily averaged fluxes[19,29]. We refer to EC data as 'ECET' throughout the article. Closed ECET data were considered to be most representative of actual ET[30]. To sample RSET pixels for comparison with ECET, flux footprints were developed for each station. Flux footprints are two-dimensional mappings of the areal extent of a station's source area, that is, the area on the ground that contributes to fluxes measured by the tower instrumentation. Refer to Methods and Volk et al.[19,20] for details on flux data processing and footprint mapping methods used. Additional discussion of uncertainty in EC data and steps taken to limit that uncertainty are provided in Supplementary Discussion 1. An overview of the satellite-driven ET models in the OpenET ensemble is provided in Methods.

The discussion of statistical results that follows focuses on comparisons between monthly aggregated ECET and RSET. Although accuracy assessments were conducted using daily (date of overpass) data and monthly total ET aggregated to growing season and annual periods, our discussion focuses on monthly results for several reasons: monthly ET has utility for longer-term water accounting and planning; uncertainties in EC data due to closure and other factors are reduced

Table 1 | Smmary statistics between modelled and observed monthly ET for cropland sites

Land cover type	Statistic	Ensemble	DisALEXI	eeMETRIC	geeSEBAL	PT-JPL	SIMS	SSEBop	*N* sites	*N* data points
All crops, mean station ET of 91 (mm per month)	Slope	0.92	0.92	0.95	0.85	0.91	0.99	0.95	53	1,652
	MBE (mm)	−5.27 (−5.8%)	−7.72 (−8.4%)	−2.44 (−2.7%)	−12.18 (−13.3%)	−2.9 (−3.2%)	4.32 (4.7%)	−6.08 (−6.7%)	44	1,638
	MAE (mm)	15.84 (17.3%)	19.91 (21.8%)	21.23 (23.2%)	22.69 (24.8%)	18.12 (19.8%)	17.93 (19.6%)	22.4 (24.5%)	44	1,638
	RMSE (mm)	20.44 (22.4%)	25.35 (27.7%)	26.97 (29.5%)	29.05 (31.8%)	23.67 (25.9%)	23.1 (25.3%)	27.72 (30.3%)	44	1,638
	r^2	0.9	0.86	0.83	0.83	0.87	0.86	0.85	53	1,652
Annual crops, mean station ET of 85 (mm per month)	Slope	0.93	0.92	0.98	0.85	0.9	1.01	0.92	42	1,446
	MBE (mm)	−5.11 (−6.0%)	−8.18 (−9.6%)	0.23 (0.3%)	−12.38 (−14.6%)	−3.77 (−4.4%)	6.27 (7.4%)	−9.13 (−10.7%)	36	1,436
	MAE (mm)	15.26 (17.9%)	20.09 (23.6%)	20.44 (24.0%)	22.52 (26.5%)	17.0 (20.0%)	17.48 (20.5%)	21.93 (25.8%)	36	1,436
	RMSE (mm)	19.71 (23.2%)	25.68 (30.2%)	26.17 (30.8%)	28.67 (33.7%)	22.31 (26.2%)	22.49 (26.4%)	27.14 (31.9%)	36	1,436
	r^2	0.9	0.84	0.83	0.82	0.87	0.85	0.84	42	1,446
Orchards, mean station ET of 126 (mm per month)	Slope	0.87	0.88	0.81	0.84	0.88	0.93	0.97	5	141
	MBE (mm)	−11.9 (−9.4%)	−11.02 (−8.7%)	−20.66 (−16.4%)	−15.11 (−12.0%)	−7.39 (−5.8%)	−3.47 (−2.7%)	−3.69 (−2.9%)	5	141
	MAE (mm)	21.18 (16.8%)	22.2 (17.6%)	28.19 (22.3%)	24.9 (19.7%)	24.43 (19.3%)	22.95 (18.2%)	20.18 (16.0%)	5	141
	RMSE (mm)	27.89 (22.1%)	27.86 (22.1%)	35.26 (27.9%)	32.77 (25.9%)	31.67 (25.1%)	30.49 (24.1%)	27.26 (21.6%)	5	141
	r^2	0.91	0.89	0.89	0.88	0.88	0.86	0.89	5	141
Vineyards, mean station ET of 112 (mm per month)	Slope	1.02	1.02	0.95	0.95	1.09	0.92	1.25	3	61
	MBE (mm)	5.27 (4.7%)	4.99 (4.5%)	−4.62 (−4.1%)	−3.76 (−3.4%)	17.95 (16.0%)	−7.71 (−6.9%)	31.88 (28.5%)	3	61
	MAE (mm)	13.66 (12.2%)	13.01 (11.6%)	18.73 (16.7%)	20.72 (18.5%)	21.53 (19.2%)	14.43 (12.9%)	33.22 (29.7%)	3	61
	RMSE (mm)	16.23 (14.5%)	15.87 (14.2%)	22.1 (19.7%)	27.2 (24.3%)	27.19 (24.3%)	17.34 (15.5%)	36.56 (32.6%)	3	61
	r^2	0.9	0.9	0.81	0.73	0.83	0.88	0.84	3	61

Mean monthly summary statistics for comparisons between OpenET[5] ensemble members' ET and closed flux tower monthly ET[19,20], grouped by three major crop types: annual crops; vineyards and orchards. Slope is calculated as the linear regression slope forced through the origin. Measures of MBE, MAE and RMSE include the error in mm per month and normalized as a percentage of the weighted mean closed flux tower ET. Note, there were three additional vegetable crop sites included in the combined crop group, which alone did not meet our data requirements for statistical analyses[19].

at the monthly (compared with daily) timescale, and OpenET directly provides daily and monthly ET, along with data services that allow users to compute ET at other aggregation periods. Accuracy results are provided for daily, monthly, seasonal and annual timesteps in Supplementary Tables 2–6, and accuracy metrics for daily timesteps should be consulted for applications of ET data at timesteps of 1–15 days. Five well-known statistical metrics were used to evaluate OpenET accuracy (for equations, see Methods): the linear regression slope forced through the origin which measures bias (Slope), mean bias error (MBE), mean absolute error (MAE), root-mean-square error (RMSE) and the coefficient of determination (r^2). Regression results with a non-zero intercept for monthly data are provided in Supplementary Table 7.

Performance over all agricultural flux sites

Of all the general land cover types sampled, OpenET models showed the strongest agreement with ECET collected in agricultural settings. For 44 agricultural sites combined, eeMETRIC, SIMS and PT-JPL showed the least bias in terms of MBE, all less than −4.5 mm per month or 5% of the mean ECET (Table 1). The ensemble value had a slightly higher magnitude bias of −5.3 mm per month, or 5.8% of the mean ECET. The ensemble value outperformed each individual model in terms of MAE 15.9 mm per month (17.3% of the mean ECET), RMSE 20.4 mm per month (22.4%) and r^2 (0.90). In comparison, MAE from individual models ranged from 17.9 to 22.7 mm, RMSE from 23.1 to 29.1 mm per month and r^2 from 0.83 to 0.87 with smallest errors from PT-JPL, SIMS and DisALEXI.

ET data from the individual RSET models were generally linearly related to ECET, with PT-JPL and SIMS exhibiting some curvature due to seasonally varying biases (Fig. 2). Many of the models underestimated ET during the cold season relative to the ECET, leading to the slightly low bias in the ensemble ET value (Table 2). To investigate seasonal variability in model accuracy, we pooled all monthly paired (model–measured) ET to generate monthly climatologies for major land cover classifications (Fig. 3 and Extended Data Figs. 1–5). The range between unclosed and closed ECET provides one measure of the uncertainty in the in situ data[31].

For most months, the multi-model ensemble ET value was well bounded between the closed and unclosed mean ECET for cropland sites, while individual ensemble members showed more seasonal bias. In spring, SSEBop and eeMETRIC underestimated unclosed ET, whereas SIMS overestimated closed ET, probably due to the assumption of well-watered conditions. In peak summer months, most models were in good agreement with closed ECET, with geeSEBAL and PT-JPL biased low. In September and October, when actual ET rates decline quickly, several models were biased high, except DisALEXI and geeSEBAL, which tracked closer to the unclosed values. The higher agreement of RSET with ECET during the peak summer period is encouraging, as this is the period of intensive irrigation and consumptive use of water through ET. A post hoc test showed that DisALEXI, geeSEBAL and SSEBop had mean monthly ET values that were statistically different (as underestimation) from the mean closed ECET. The mean aggregated growing season ET for all models were no different from the mean closed ECET (Supplementary Tables 8 and 9).

The monthly climatologies derived at flux sites were upscaled using data from all cropland pixels over the full OpenET domain (Extended Data Fig. 6). We found similar seasonal patterns and relative model biases to those identified at the flux sites—giving confidence in the representativeness of the ECET comparisons.

Impact of sampling interval on model performance

Model accuracy often improves with temporal aggregation interval due to cancellation of errors[8]. In croplands, the accuracy metrics for the OpenET ensemble improved as the aggregation period increased

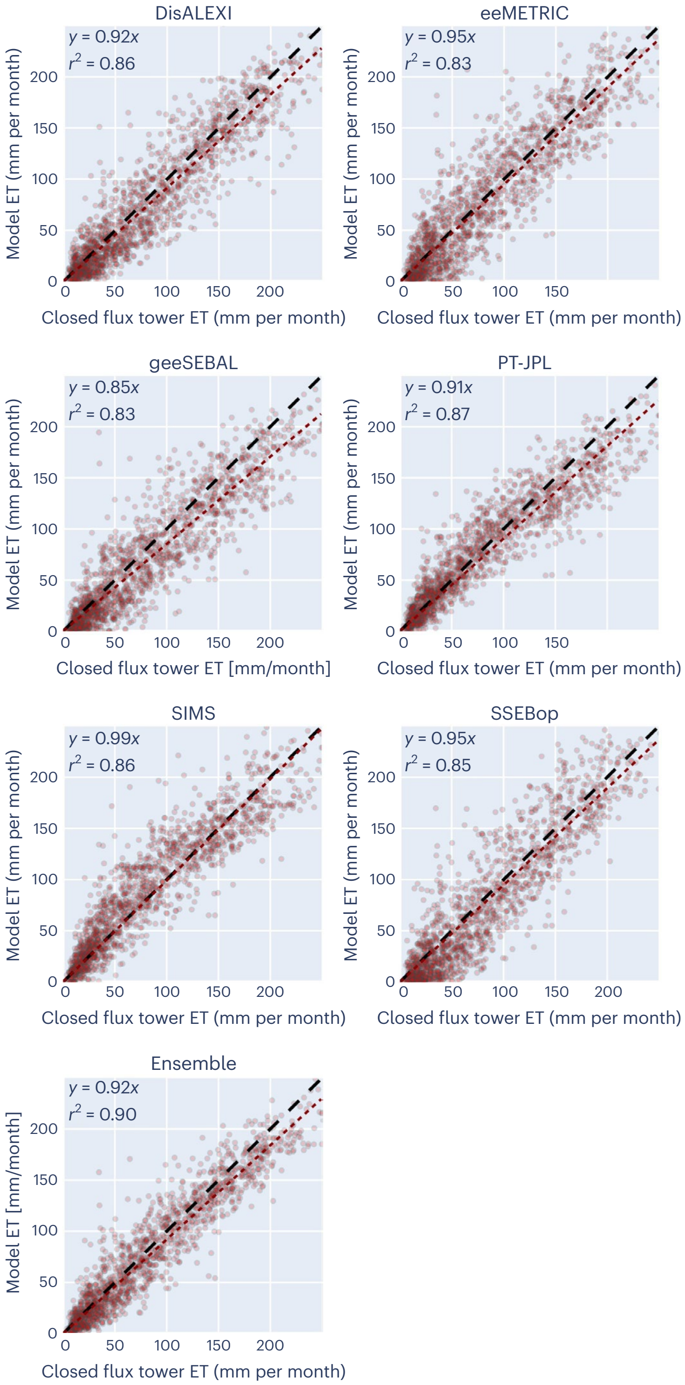


Fig. 2 | Modelled versus observed monthly ET at cropland sites. Monthly comparison of all paired OpenET[5] ensemble members ET versus closed flux tower ET[19,20] from all cropland stations for all months of record. Included for each model is the result of the least square linear regression model forced through the origin and r^2.

from daily (overpass dates) to monthly to growing season to annual periods (Supplementary Tables 2–6). Daily ensemble results for the combined cropland sites showed a MAE of 23.6%, and RMSE of 31.1% of the mean ECET. At this timescale there is increased uncertainty both in the ECET data due to variability in micrometeorological conditions and energy balance closure, and remotely sensed ET due to potential cloud contamination and errors in footprint representation. These ensemble uncertainties are reduced when integrating to monthly (MAE of 17.3% and RMSE of 22.4% of ECET), growing season (MAE of 12.9% and RMSE of 15.5% of ECET) and water year (MAE of 11.3% and RMSE of 12.3% of ECET) timescales. Fortunately, during growing season periods we found lower energy balance closure error in EC data[19] and there is less cloud cover in satellite data in the western United States as compared with the non-growing period. During the summer, the daily ensemble normalized MAE (NMAE) on overpass dates was typically between 5% and 25% (Supplementary Fig. 1), and monthly 7% and 20% (Fig. 4). We expect custom aggregation periods between 2 and 15 days to have similar or slightly improved accuracy to daily results that vary seasonally; subweekly to bi-weekly RSET may be of greatest use for irrigation scheduling[32].

Performance among annual and perennial crops

Annual crops, including wheat, corn, soy, rice and others, make up the majority (80%) of cropland sites in the OpenET ECET dataset (Supplementary Table 1). Compared with perennial crops, annual crops tend to have shorter canopies and more homogeneous cover at peak growth stage. The annual crop sites in the OpenET flux dataset are predominantly irrigated, and are distributed across a range of climatic zones, with higher density in regions such as Mediterranean and semi-arid Central Valley, California, and humid continental regions in the High Plains and the Mississippi Alluvial Plain (Fig. 1).

For annual crops, each of the RSET models in the OpenET ensemble exhibited small bias and high levels of accuracy and precision (Table 1). Similar to all crop types combined, the ensemble value for annual crops outperformed individual models in terms of MAE (15.3 mm per month or 17.9% of mean ECET), RMSE (19.7 mm per month or 23.2% of mean ECET) and r^2 (0.9). Of the RSET models, eeMETRIC and PT-JPL exhibited the lowest magnitude of MBE, with PT-JPL and SIMS yielding the highest accuracy in terms of MAE and RMSE.

Dividing annual crops into C3 and C4 subclasses, we find the seasonal patterns and magnitudes of ensemble MAE are similar throughout the year (Fig. 4). NMAE in general reflects the inverse of the characteristic water use curve for each class, with C3 crops exhibiting a broader seasonal curve than C4 and therefore lower NMAE early and late in the season. While the higher NMAE values observed outside the growing season for all crop types (Fig. 4) are more indicative of low ET rates than of meaningful modelling error characteristics, cool-season errors may be generally inflated by higher cloud cover, increasing the time interval between cloud-free satellite retrievals. Improving satellite imaging frequency, as well as ET time integration and gap-filling techniques, should help to increase OpenET accuracy during the non-growing season (Discussion).

Another class of interest is woody perennials, which are high-value crops and pose distinct modelling challenges. High-quality eddy flux ET data were available for three vineyards, three nut tree orchards and one fruit orchard, all located in California[19,33]. Vineyards and orchards have taller and more highly structured canopies, often with inter-row cover crops, and vineyards are often deficit irrigated. These qualities lead to shadowing and mixed pixel effects in remote sensing at the 30-m level, and the need for sensitivity to small changes in vine stress to inform deficit irrigation applications is a unique modelling requirement.

RSET model performance in the vineyard sites sampled was strong and consistent across models. The ensemble accuracy exceeded that for annual crops (Table 1 and Fig. 4), with lower bias (slope of 1.02 and MBE of 5.3 mm per month) and lower MAE and RMSE (13.7 and 16.2 mm per month, respectively, or 12.2% and 14.5% of the mean monthly ECET) and r^2 of 0.90. DisALEXI performed similarly or better than the ensemble at the vineyard flux sites, perhaps due to its two-source approach towards partitioning temperature fluxes between the substrate (inter-row) and canopy.

Performance was more varied across ensemble members for the orchards than for other broad crop types, and biases were more negative. This could be related to shadowing effects in the taller and more

Table 2 | Summary statistics between modelled and observed monthly ET for cropland sites grouped by climate zone

Land cover type	Statistic	Ensemble	DisALEXI	eeMETRIC	geeSEBAL	PT-JPL	SIMS	SSEBop	*N* sites	*N* data points
Bsk + Bsh (cold and hot semi-arid steppe), mean station ET of 133 (mm per month)	Slope	0.9	0.85	0.91	0.82	0.88	0.97	1	11	246
	MBE (mm)	−6.94 (−5.2%)	−15.17 (−11.4%)	−4.65 (−3.5%)	−18.09 (−13.6%)	−7.55 (−5.7%)	2.71 (2.0%)	2.89 (2.2%)	11	246
	MAE (mm)	20.74 (15.6%)	26.93 (20.3%)	26.94 (20.3%)	30.31 (22.8%)	25.7 (19.3%)	22.99 (17.3%)	24.23 (18.2%)	11	246
	RMSE (mm)	26.38 (19.8%)	34.55 (26.0%)	33.4 (25.1%)	38.0 (28.6%)	32.49 (24.4%)	28.97 (21.8%)	30.99 (23.3%)	11	246
	r^2	0.89	0.81	0.8	0.8	0.84	0.84	0.84	11	246
Bwh + Bwk (hot and cold desert), mean station ET of 110 (mm per month)	Slope	0.91	0.85	1.02	0.92	0.86	0.92	0.88	10	53
	MBE (mm)	−6.78 (−6.1%)	−15.63 (−14.2%)	9.63 (8.7%)	−8.19 (−7.4%)	−7.77 (−7.0%)	−3.2 (−2.9%)	−12.34 (−11.2%)	7	49
	MAE (mm)	13.24 (12.0%)	21.21 (19.2%)	18.92 (17.1%)	19.13 (17.3%)	19.21 (17.4%)	13.62 (12.3%)	19.59 (17.8%)	7	49
	RMSE (mm)	17.02 (15.4%)	25.78 (23.4%)	23.92 (21.7%)	23.51 (21.3%)	23.81 (21.6%)	16.07 (14.6%)	22.95 (20.8%)	7	49
	r^2	0.91	0.85	0.88	0.87	0.83	0.94	0.89	10	53
Cfa (humid subtropical), mean station ET of 75 (mm per month)	Slope	1	1.03	1.03	0.99	0.93	1.15	0.91	11	232
	MBE (mm)	2.15 (2.9%)	4.49 (6.0%)	3.65 (4.9%)	0.97 (1.3%)	−0.88 (−1.2%)	14.98 (19.9%)	−3.84 (−5.1%)	8	228
	MAE (mm)	17.51 (23.3%)	20.17 (26.8%)	24.01 (31.9%)	20.06 (26.7%)	17.79 (23.6%)	22.71 (30.2%)	21.97 (29.2%)	8	228
	RMSE (mm)	23.76 (31.6%)	26.18 (34.8%)	31.88 (42.4%)	28.39 (37.7%)	23.62 (31.4%)	30.23 (40.2%)	28.76 (38.2%)	8	228
	r^2	0.75	0.72	0.62	0.69	0.76	0.72	0.64	11	232
Csa + Csb (hot- and warm-summer Mediterranean), mean station ET of 94 (mm per month)	Slope	0.95	0.96	0.99	0.81	1.01	0.93	0.99	8	292
	MBE (mm)	−4.37 (−4.7%)	−6.32 (−6.7%)	−1.72 (−1.8%)	−20.92 (−22.3%)	7.37 (7.9%)	−1.34 (−1.4%)	−2.9 (−3.1%)	8	292
	MAE (mm)	13.32 (14.2%)	18.65 (19.9%)	17.14 (18.3%)	25.9 (27.6%)	17.17 (18.3%)	14.04 (15.0%)	23.94 (25.6%)	8	292
	RMSE (mm)	16.52 (17.6%)	22.75 (24.3%)	21.48 (22.9%)	31.31 (33.4%)	21.97 (23.5%)	18.27 (19.5%)	27.92 (29.8%)	8	292
	r^2	0.93	0.87	0.88	0.85	0.87	0.88	0.84	8	292
Dfa + Dfb (hot- and warm-summer humid continental), mean station ET of 67 (mm per month)	Slope	0.9	0.93	0.94	0.86	0.87	1.02	0.9	13	829
	MBE (mm)	−8.0 (−11.9%)	−8.07 (−12.0%)	−6.74 (−10.0%)	−10.55 (−15.7%)	−5.72 (−8.5%)	4.88 (7.3%)	−13.58 (−20.2%)	10	823
	MAE (mm)	13.8 (20.5%)	15.68 (23.3%)	18.93 (28.2%)	17.88 (26.6%)	13.67 (20.3%)	15.35 (22.8%)	21.09 (31.4%)	10	823
	RMSE (mm)	17.85 (26.6%)	20.36 (30.3%)	24.1 (35.8%)	23.36 (34.7%)	18.89 (28.1%)	19.94 (29.7%)	25.9 (38.5%)	10	823
	r^2	0.91	0.9	0.83	0.86	0.89	0.86	0.86	13	829

Mean monthly summary statistics for comparisons between OpenET[5] ensemble members ET and closed flux tower monthly ET[19,20] for agricultural sites grouped by Köppen–Geiger climate zones[34]. Slope is calculated as the linear regression slope forced through the origin. Measures of MBE, MAE and RMSE include the error in mm per month and normalized as a percentage of the weighted mean closed flux tower ET.

strongly clumped canopies, particularly for models that are strongly dependent on LST inputs. The ensemble value had a negative bias with mean slope of 0.87, MBE −11.9 mm per month, MAE 21.2 mm per month (16.8% of ECET) and RMSE 27.9 mm per month (22.1% of ECET), and an r^2 of 0.91. SSEBop and SIMS had the least bias in terms of slope and MBE, and SSEBop and DisALEXI had the lowest error in terms of MAE and RMSE (Table 1). While MAE in orchards is high mid-season, the normalized values are similar to those of annual crops (Fig. 4).

Variation of model performance across climate regions

To investigate variations in OpenET performance over different climates, cropland accuracy metrics were grouped by the Köppen–Geiger climate zones of the flux sites[34] (Fig. 1). Zones with fewer than five flux stations were omitted as a conservative measure, and some zones were lumped on the basis of secondary climate classifications (for example, hot- and warm-summer Mediterranean zones). Each resulting group had 7–13 flux stations used for calculation of accuracy statistics.

Overall, the OpenET ensemble had better agreement with ECET at crop sites in water-scarce, semi-arid to arid regions (Mediterranean and desert zones in the Southwest) as compared with humid zones (Table 2 and Supplementary Fig. 2). Irrigation is more prevalent in semi-arid to arid regions, and crop ET tends to be closer to potential ET rates and is more accurately modelled in some RSET modelling frameworks. High accuracy of models in semi-arid and arid regions is advantageous, given the high priority of water resource sustainability and management challenges in these regions.

Among the zones considered, the OpenET ensemble value was most accurate for crop sites in Mediterranean zones, with MAE of 13.3 and RMSE of 16.5 mm per month (14.2% and 17.6% of the mean ECET), with the ensemble outperforming individual members. Of the individual models, SIMS showed the best agreement with ECET in these regions, suggesting well-watered conditions for most sites or possible influence of adjacent non-irrigated areas on SEB models. Similarly, in arid sites (hot and cold desert), SIMS had the lowest MAE and RMSE (Table 2). During the growing season periods when the majority of irrigation is applied, the ensemble's monthly NMAE was consistently below 10% for cropland sites in Mediterranean climates (Supplementary Fig. 2).

Model performance in the subhumid and humid continental regions of the Midwest and Central Plains was similar to that in the Mediterranean climate zone, again with the ensemble outperforming individual models in terms of collective statistics (Table 2 and Supplementary Fig. 2). Errors were higher at the humid subtropical sites, with SIMS tending to overestimate ET with a slope of 1.15 and normalized MBE of 19.9%, indicating ET is less well correlated with vegetation density in this region, and that irrigation practices may result in intermittent vegetation water stress. Hypotheses for increased RSET error in humid regions and paths for improvement are proposed in Discussion.

Performance in natural ecosystems

Most of the flux stations (61%) used in the intercomparison were in non-agricultural sites, including shrublands, grasslands, mixed forests, conifer forests, and wetlands or riparian areas (Fig. 1)[19]. The SIMS model is currently not designed for and implemented in non-agricultural land-cover types; for these pixels, the ensemble consists of five models with the possibility of removing a single outlier (Methods). Systematic model error and variability for non-agricultural sites was higher than cropland sites (Fig. 5).

Most models exhibited a high bias in wetland/riparian sites, dominated by overprediction of ET during the spring (Extended Data Fig. 5). SSEBop had higher accuracy in these sites than other models and the ensemble value (Supplementary Tables 2–4). For models that estimate all components of the SEB (DisALEXI, eeMETRIC and geeSEBAL), this bias could result from an underestimation of the substrate (water) heat storage term in the spring before the vegetation canopy develops[7]. These errors can potentially be mitigated in the future through accurate classification of inundated land areas.

Natural ecosystems under high water stress, such as shrublands and grasslands in desert and semi-arid steppe climates in the western United States, showed the highest variability and error with respect to ECET (Fig. 5 and Supplementary Tables 2–4). In these systems, ET can be a small fraction of available energy, and difficult to both measure on the ground and model using RSET approaches. Shrublands also tend to be more heterogeneous than cropland sites, and this can introduce additional uncertainty into model–measurement comparisons[5]. Nevertheless, it is important to provide an evaluation of accuracy, both to benefit ET monitoring and land health assessments within shrub and grassland ecosystems, and to identify key areas for future research in RSET to reduce model error.

The Landsat-scale ET from OpenET also has applications in forested landscapes, as a predictor of forest health and mortality[35] and as a metric of water yield response to forest management[36]. In forested locations, most OpenET models overestimated ET, particularly at the evergreen flux sites sampled, yielding a slope for the ensemble value of 1.24 and MBE of 16.8 mm per month (27.3%). At these sites, eeMETRIC showed the least bias with a slope of 1.17 and an MBE of 10.8 mm per month (17.5%), while for MAE and RMSE, the ensemble value outperformed each individual model. At mixed forest sites, however, eeMETRIC and DisALEXI were in better agreement with ECET than was the ensemble.

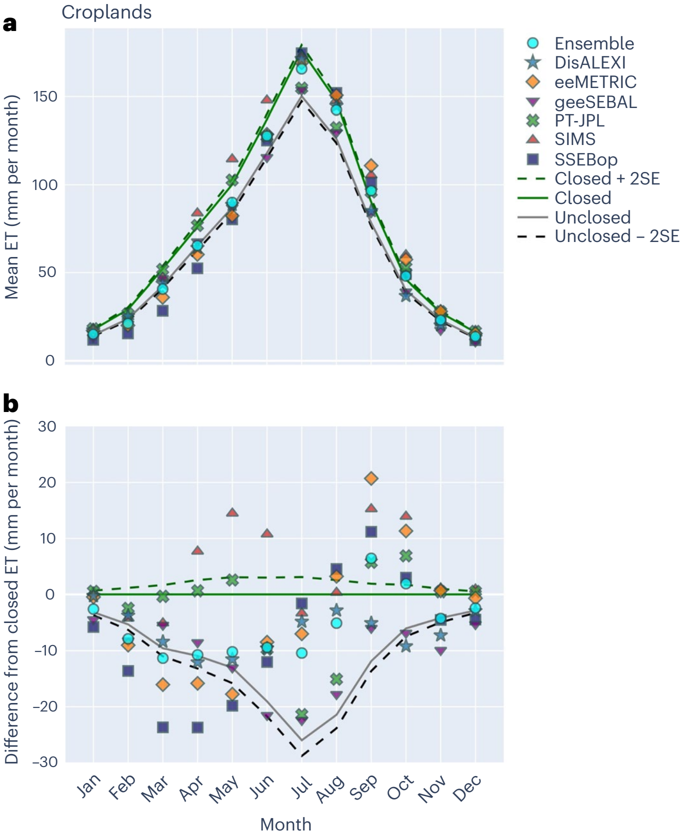


Fig. 3 | Monthly climatology of paired modelled and observed ET for cropland sites. a, Monthly climatology of paired OpenET[5] and flux tower ET[19,20] from cropland sites. **b**, The residual of monthly mean ET (model minus mean closed flux ET). Unclosed and closed labels refer to flux tower ET before and after energy balance closure correction. Dashed lines represent the closed flux ET mean plus two standard errors of the mean and unclosed flux ET mean minus two standard errors of the mean.

Ensemble outlier removal and spatial inter-model variability

See Supplementary Discussion 2 for analysis and discussion of the MAD outlier removal approach that is used for computing the ensemble value, including spatial analysis of the occurrence of outliers and the long-term differences between each model's seasonal ET and the ensemble value (Extended Data Figs. 7 and 8, Supplementary Figs. 3–9 and Supplementary Tables 9 and 10). Evidence suggests that the MAD approach showed accuracy metrics similar to other simple methods. Over 2016–2022, typically no model was identified as an outlier in cropland pixels; however, SIMS was about 10% more likely to be identified as an ensemble outlier, and it often gave the highest ET value, particularly in the Central Plains.

Discussion

ET is a critical driver and metric of ecosystem function, weather and climate, agricultural practices and water resource management. However, field-scale ET has previously been difficult to estimate at scale; therefore, ready access to high-resolution (spatially and temporally) ET data offers societal benefits to a variety of stakeholders[1,5]. Using monthly ET data, water managers can develop more accurate water budgets in support of incentive-driven conservation programmes and innovative management and trading strategies. For policymakers, such data can improve water supply tracking, simplify regulatory compliance and promote the co-development of solutions with local communities. Crop producers may be able to improve the efficiency of irrigation practices in some instances, resulting in enhanced sustainability and reduced costs for water, fertilizer and energy. Supplementary Discussion 3 continues the conversation on incentives towards improving irrigation efficiency and how OpenET data can provide value in an RSET-based irrigation scheduling framework.

In addition to informing water management, OpenET has multiple research and modelling applications. Carbon and climate modelling can benefit from 30-m RSET data as a diagnostic indicator of ecosystem health and function response under a changing climate[1]. RSET is being used to reduce summertime warm-dry bias in weather forecasting and climate models by improving the representation of ET from irrigated land[37], ET–soil moisture coupling[38] and transpiration–evaporation partitioning[39]. Hydrologic and land surface models at multiple scales can also benefit from high-resolution ET data, for example, as validation or forcing data in basins where streamflow measurements are not available to constrain the water budget[13,40,41].

Realizing the full potential benefits of RSET data for water resource and land management applications requires rigorous and reproducible accuracy assessment to inform practitioners on best use practices[18]. The accuracy results we present here provide valuable constraints on model uncertainty based on broad crop type, climate region and timescale.

Average error in the OpenET ensemble value with respect to mean ECET in cropland sites for monthly, growing season and annual aggregated ET, ranged from 10% to 17% for MAE and 11% to 22% for RMSE. These errors are within accuracy levels of 10–20% reported for supervised remote sensing techniques[42]. They are also consistent with accuracy targets set by the OpenET user groups: 10–20% at a

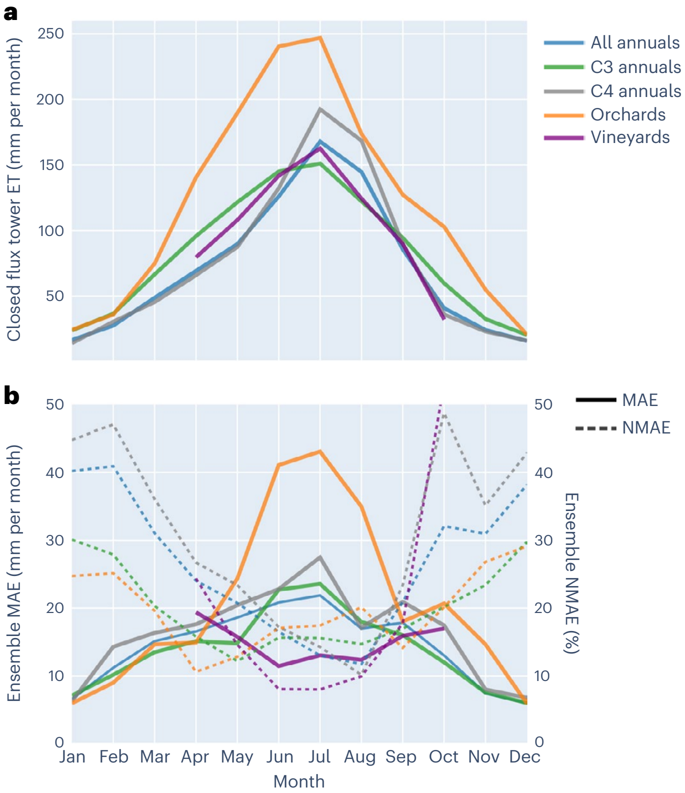


Fig. 4 | Monthly MAE of the model ensemble for different crop types. **a**, Monthly mean flux tower ET[19,20]. **b**, OpenET[5] ensemble MAE and MAE normalized by the mean flux tower ET[19,20] (NMAE) using all paired model–measured data for cropland stations grouped by crop types. Annual crops that had a mixed history of rotation between C3 and C4 crop types, for example, corn–soy rotations, were not included in C3 or C4 results but were included in the combined grouping.

monthly timestep, and 15–25% for daily ET data[5]. These errors include uncertainties in ECET data, which are estimated to range from 10% to 30% depending on site characteristics and instrumentation design and maintenance[42].

These accuracy results may support advancements in water management applications that incorporate OpenET data. For croplands, all models except for SIMS had negative bias errors at the monthly timestep (−2.7% to −13.3%), with an MBE of −5.8% for the ensemble ET value (SIMS MBE is +4.7%). Awareness of these bias errors when using these data for irrigation management applications may prevent unintentional deficit irrigation that can suppress crop yields and farm revenue[43]. Cross-comparisons between the primarily reflectance-based SIMS and PT-JPL models and the LST-driven models may be useful for identifying periods of intentional or unintentional crop water stress and deficit irrigation. Reducing errors in the OpenET daily data is a high priority for advancing their utility for on-farm water management.

At local to regional scales, the reported uncertainties at monthly to annual timesteps should inform applications related to water balance, water accounting and water rights administration. Comparison of OpenET data aggregated at the scale of irrigation districts or watersheds against carefully constrained water balances offers one path to assessment of biases at larger scales. Particularly in administration of water rights, the current uncertainty in the OpenET data (for example, growing season ensemble NMAE of 12.9% for croplands) must be recognized in evaluating consumptive water use, and OpenET data should only be used for this purpose in combination with other sources of information.

This study provides insights into potential pathways towards improving the accuracy of the individual models within the OpenET ensemble. Across both agricultural and some natural landscapes, most models underestimated cropland ET during the winter and spring, particularly the models that rely upon TIR measurements to compute ET. This underestimation may be related to loss of thermal contrast over an image, where differences between the hottest and coolest pixels are reduced relative to midsummer values, adding uncertainty to within-scene scaling approaches. It may also be related to misrepresentation of soil evaporation during extended wet periods, extended periods of cloudiness, and error in shared model inputs. In addition, treatment of effects of senesced standing vegetation and crop residue on SEB can impact model performance outside of the growing season. In terms of observational errors, the energy balance closure error and uncertainty in EC data are also amplified during periods outside of the growing season[19].

We found increased model error in croplands in humid climates as compared with drier regions. Again, lower temperature contrasts across humid landscapes may contribute to errors in TIR-based within-scene scaling models. A primary driver, however, is probably the relative paucity of clear-sky satellite retrievals and potential for error in LST due to undetected clouds. Improving temporal sampling of RSET model inputs will be a major focus of on-going development in OpenET, through future use of imagery from additional Landsat-like optical (Sentinel-2) and thermal (ECOSTRESS, VIIRS) sensors[44], and integration of future TIR observations from satellite missions currently in development by NASA, USGS and the European Space Agency. Methods for computing ET values between cloud-free satellite observations, currently based on linear interpolation of the ratio of ET to a reference flux, can also be improved. Approaches used in mapping and predicting vegetation phenology[45] and dynamic time warping[46] algorithms developed for signal processing applications offer promise for reducing large errors during periods of rapid vegetation change or extended cloud cover, which would contribute to reduced RMSE values across the model ensemble.

Examining results for specific crop classes, we found strong results for DisALEXI and SIMS over vineyards, and DisALEXI, SIMS and SSEBop over fruit and nut orchard sites—key targets for irrigation management in the Central Valley. Increasing the number of validation sites in orchards would help to address remaining modelling issues associated with this challenging canopy architecture. The USDA ARS-led Tree-crop Remote sensing of Evapotranspiration eXperiment (T-REX) is aimed at addressing this observational gap[47].

All models, to varying degrees, have room for notable improvement in computation of ET in natural ecosystems. For example, most models systematically underestimate ET in drier ecosystems such as grasslands and shrublands and overestimate ET in evergreen forests. Incorporation of high-frequency and high-resolution visible and near-infrared data into the remote sensing models may improve their ability to capture phenological shifts particularly in arid/semi-arid regions, and agricultural systems in general[48,49]. Improvement of gridded meteorological model inputs[50,51], land cover classification data and soils data[52] may also lead to improved model performance in both natural ecosystems and in croplands. In particular, datasets compiled from agricultural weather stations and used to compute bias correction surfaces for reference ET could be re-evaluated to ensure reference surface compliance with the assumptions of the American Society of Civil Engineers Penman–Monteith equation[53].

Future OpenET accuracy evaluations will target primary causes of error in ground ET measurements and RSET methods. Specific factors to consider include local advective impacts on modelled and measured ET, EC energy budget closure, local thermal contrast, ET reduction in deficit irrigated or rainfed systems, potential biases in gridded meteorological inputs to RSET models, and accurate capture of ET over sparsely cultivated landscapes. Comparisons with other

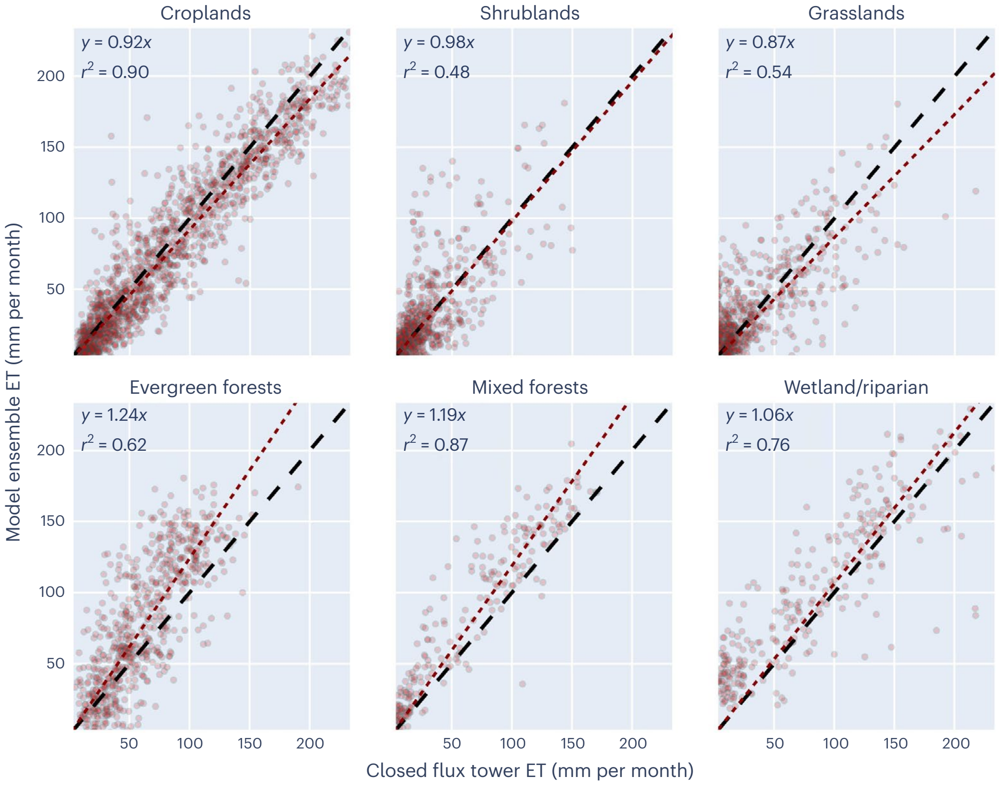


Fig. 5 | Monthly modelled ensemble versus observed ET for sites grouped by land cover type. Monthly comparison of the paired monthly OpenET[5] ensemble ET versus closed flux tower ET[19,20] for each general land cover group. Included for each group is the result of the least square linear regression model and r^2.

well-established spatially mapped ET products such as MOD16 or FLUXCOM[54] may provide further insights for operational global ET mapping at field scales (30–100 m). Comparisons against ET data computed from long-term water balance studies[13,55] would help fill in gaps of spatial coverage in measured in situ ET across the western United States in hydrologically important but sparsely cultivated regions such as the Upper Colorado River Basin.

Conclusions

The OpenET platform provides spatially continuous ET data at 30-m resolution throughout the western United States. An intercomparison and accuracy assessment involved six satellite-based RSET models composing the current OpenET version, ensemble ET computed from the six models, and a well-documented benchmark eddy flux dataset from 152 stations located in the contiguous United States. Based on results from 59 cropland ET stations located in a variety of climatic regions, little systematic model bias was observed in croplands, and error metrics were within or near the targets set forth by OpenET partners including farmers, irrigation managers and water management agencies. The best accuracy metrics were associated with seasonal and annual timescales, and for crops in arid/semi-arid regions. The OpenET ensemble mean, with outlier removal, typically outperformed any individual model in terms of error statistics. Generally, no more than one model was identified as an outlier during growing season months over most agricultural regions in the western United States, and frequently no models were excluded. This finding highlights the substantial progress achieved so far in developing fully automated RSET modelling approaches that can be employed to map ET over large areas at field-scale resolution. The study identified paths for future targeted research and model improvement, and is intended to support the RSET research community in the development of increasingly robust and accurate RSET techniques. We are also hopeful that this assessment will provide added confidence to water resource managers, farmers, ranchers, scientists and other potential users of OpenET due to the high rigour and transparency of methods that were employed.

Methods

Flux data processing and footprint sampling

We used a curated benchmark eddy flux-based ET dataset[19,20] and tools[56] for use in this and subsequent evaluations of OpenET RSET models[5]. The rationale and decision-making steps for the collection and post-processing of flux data, as well as analyses of footprint sampling techniques and energy balance closure error within the dataset, are described in Volk et al.[19,20]. Data processing techniques for gap-filling and correction for energy balance closure error were conducted using open-source Python tools[56] that enhance data provenance and reproducibility. Data were also subject to qualitative, visual-based data screening and filtering[19,20]. The final post-processed dataset consists of 161 stations, is public and includes daily and monthly ET and meteorological data, interactive graphics of such data for each station, and site information such as land use and Principal Investigator acknowledgements[20]. We note that nine stations in the dataset were not included in the statistical results presented here because they had data coverage that did not overlap with the data that could be developed for all six OpenET models. For example, not all models could be implemented from satellite imagery recorded before 2001 (ref. 5). Figure 1 shows a map of the 152 stations used in this accuracy assessment as well as their land cover types and Köppen–Geiger climate zones, and Supplementary Table 1 provides additional metadata for each station.

Data for the majority (106) of the flux stations in this study were downloaded from the AmeriFlux website, last accessed on 27 October 2020, and the remaining stations were retrieved from a variety of sources and Principal Investigators from university partners, the US Geological Survey, the US Department of Agriculture and others[19]. In addition to EC systems, four precision weighing lysimeters measuring cropland ET in Texas[57] and seven high-quality Bowen Ratio instrumented sites, which measure ET in predominantly phreatophyte shrublands in Nevada[20], were included in the dataset. Gap-filling of initial half-hourly fluxes of the four main energy balance components—latent, sensible and soil heat flux, and net radiation—was conducted using linear interpolation where gaps up to 2 h during the daytime or 4 h during nighttime were interpolated. If a given 24-h period still contained gaps then the daily average was not calculated and the daily flux value was left as a gap. After this initial gap-filling, fluxes were averaged to daily periods and energy balance closure correction was applied following the daily energy balance ratio approach defined by FLUXNET2015/ONEFlux[19,29]. The corrected daily latent heat flux, which is the energy consumed through ET, was used to calculate ET with an adjustment to the latent heat of vapourization for air temperature[20]. This closure-adjusted value is referred to as closed flux ET or measured ET in the main text and all statistical measures reported for OpenET models were against the energy balance corrected ET data. Daily ET gaps were subsequently filled using gridMET fraction of reference ET and gridMET grass reference ET[19,20,58]. To exclude flux stations with higher data uncertainty, only stations with mean daily energy balance closure of 0.75 or higher during the growing season and 0.6 or higher during the non-growing season were chosen for this intercomparison. Here, growing season periods were spatially mapped on the basis of a cumulative growing-degree-day and killing frost approach derived from long-term gridded climate data and are specific to each flux site[19,58]. The final dataset is similar to the recent FLUXNET2015 (ref. 29) release consisting of high-quality eddy flux station data that were subject to similar processing and correction techniques. The largest difference between the two datasets, in terms of daily latent heat flux estimates, results from different gap-filling procedures, where our approach is considered to be simpler and more conservative[19,20,29].

Two approaches were used to estimate flux tower footprints or source area for tower pixel sampling of RSET imagery: (1) simple square ‘static’ pixel (Landsat 30 m) grids of 3 × 3, 5 × 5 and 7 × 7 drawn around station locations, and (2) two-dimensional, physically based flux source area estimations modelled using hourly meteorological data using the Kljun et al.[59] approach, with hourly footprints converted to daily/monthly average footprint rasters weighted by reference ET[19]. The placement of the static grids was informed by high-resolution imagery to avoid inclusion of pixels of non-representative land cover (structures, roads and canals), and shifted slightly into the predominant wind direction as determined by long-term mean daytime windroses (built from data between 6:00 and 20:00 local time). Although the physically based and temporally dynamic footprints were preferred over the static footprints, only about half of the stations in the dataset had sufficient data for their production. Commonly, one or more input parameters to the Kljun et al.[59] model, such as the standard deviation of the crosswind component of wind due to turbulence or friction velocity, was not available. A detailed description of parameter estimation, processing steps and the method used for creating weighted mean footprint images (using reference ET from NLDAS2 gridded weather data[60]) can be found in Volk et al.[19]. We also conducted a rigorous comparison of the intersection between source areas from the static grids of different sizes and the temporally dynamic footprints. The major finding was that the larger 7 × 7 grids tended to include substantially more of the dynamically defined footprint area than did the smaller grid sizes on average; however, the smaller 3 × 3 grids tended to overlap with pixels that were deemed part of the dynamic footprint on a more consistent basis. Therefore, we decided to use the 7 × 7 grids for pixel sampling at most flux sites where a dynamic footprint could not be generated, with exceptions for sites with heterogeneous surroundings or with non-representative land cover nearby the station. For these sites, we used 5 × 5 or 3 × 3 grids to avoid giving equal weight to pixels of potentially different land cover that lie near the perimeter of the typical actual footprint area[19].

Model data

The majority of the models that make up the OpenET ensemble are based on full or simplified implementations of the SEB approach. The SEB approach accounts for the energy used to transform liquid water in plants and soil into vapour that is released to the atmosphere. The SEB approach relies on satellite measurements of surface temperature and surface reflectance combined with other key land surface and weather variables to calculate components of the energy balance—net radiation, sensible heat flux, ground heat flux and latent heat flux. eeMETRIC[8], geeSEBAL[9] and DisALEXI[7] compute each component of the energy balance using optical (that is, short-wave) and thermal (that is, long-wave) data, whereas SSEBop[13] and PT-JPL[10] are simplified approaches in which certain components of the energy balance are not calculated, or are calculated using a set of simplifying assumptions. SIMS[11,12] relies on surface reflectance data, crop type information and a gridded soil water balance model to compute ET as a function of canopy density using a crop coefficient approach for agricultural lands.

The Google Earth Engine[14] Python application programming interface was used to develop a workflow for sampling OpenET RSET model data at ET flux sites. Sampling of the daily and monthly RSET model data was performed at each site using a set of static (3 × 3, 5 × 5 and/or 7 × 7) and/or dynamic flux source-area footprints. Conditions for each of the extraction methods using static footprints were as follows: (1) daily ET from eeMETRIC, SIMS and SSEBop for sites outside of California was calculated as the product of the mean daily fraction of grass reference ET (EToF) produced by the models and the mean daily bias-corrected gridMET grass reference ET (ETo) (repeated for sites within California using daily CIMIS ETo, where CIMIS is more commonly used and depended upon in California); (2) daily ET from PT-JPL, geeSEBAL, and ALEXI/DisALEXI for all sites was computed as the spatial average of daily ET pixels produced by the models; (3) monthly ET from all RSET models for sites outside of California were calculated as the product of the mean monthly EToF and the mean monthly gridMET ETo (repeated for sites within California using the monthly CIMIS ETo). The process of extrapolating instantaneous data (time of overpass) to daily ET is an internal model calculation and differs for each model, and we refer readers to the individual model documentations for details as well as Melton et al.[5]. Daily Landsat image pixels with cloud contamination are flagged on the basis of the CFMask derived indicators[61] in the pixel quality assurance band (QA_PIXEL) and those pixels are not considered. When computing monthly ET, all missing or masked daily ET pixels are computed by linearly interpolating between the nearest unmasked (cloud free) pixels in time within ±32 days.

Conditions for each of the extraction methods using dynamic footprints were as follows:

(1) daily ET from eeMETRIC, SIMS and SSEBop for sites outside of California was calculated by first multiplying the sampled daily EToF pixels produced by the models in the footprint by each daily flux footprint weight to obtain daily weighted EToF pixels, and summing all daily weighted EToF pixels to obtain mean daily weighted EToF, normalizing the mean daily weighted EToF by the sum of weights to account for times when the sum of weights did not equal 1 (for example, caused by cloud masking of pixels), and then multiplying the mean daily weighted EToF by the mean daily bias corrected gridMET ETo (replaced for sites within California using the daily CIMIS ETo);
(2) daily ET from PT-JPL, geeSEBAL and ALEXI/DisALEXI for all sites was calculated by multiplying the daily ET pixels by the daily flux footprint weights to obtain daily weighted ET pixels, summing all daily weighted ET pixels to obtain mean daily weighted ET, and then normalizing the mean daily weighted ET by the sum of weights, and
(3) monthly ET from all RSET models for sites outside of California was calculated by first multiplying the monthly EToF pixels by

the monthly flux footprint weights to obtain monthly weighted EToF pixels, summing all monthly weighted EToF pixels to obtain mean monthly weighted EToF, normalizing the mean monthly weighted EToF by the sum of weights, and then multiplying the mean monthly weighted EToF by the mean monthly bias-corrected gridMET ETo (replaced for sites within California using the monthly CIMIS ETo).

Additional processing was required after extracting the daily ET when duplicate days of data were extracted at select sites due to overlapping Landsat paths. Occasionally a site would lie within the footprints of two overlapping Landsat scenes, resulting in more than one ET value on a given overpass date. To obtain single daily ET values for the site, the daily weighted mean ET for each day was computed using the pixel count (that is, number of pixels used when deriving the respective spatial mean ET value) as the weight. ET pixel counts were occasionally less than the grid/footprint total because of the removal of poor-quality pixels (for example, cloud masking).

Ensemble computation

The ensemble mean of the six OpenET models was computed after removing up to two outlier models based on the MAD[23,24], a robust measure of spread that is suitable for small samples. The outlier removal occurs at the pixel level for each ET image generated. To identify outliers for a single scene, first the median value and the MAD from the median is computed as

$$\mathrm{MAD} = b \times \mathrm{median}\left(|X_i - \mathrm{median}\,(X)|\right),$$

where X_i is the ET value for model i and X is the full set of all six model's ET estimates. Here, b is a scalar set to 1.483, and it was derived on the basis of the assumption of normality of the sample population[62]. This approach is sometimes referred to as the MADe rule, where $e = 1.483$. The MAD value is typically scaled by 2, 2.5 or 3 on the basis of a subjective assessment of the data, which is then used to create a band around the median:

$$\mathrm{median}\,(X) \pm 2\mathrm{MAD}.$$

Model estimates that fall outside the band are deemed as outliers, and up to two outliers (those furthest from the median) are removed from the set of model estimates before taking the ensemble mean.

Due to the tendency for some OpenET models to predict zero ET or even negative ET rates in some arid regions during dry periods we modified the above approach for these scenarios. Specifically, when the ensemble median estimate is zero but at least one model predicts a positive ET rate, the ensemble mean is taken to include that value without any prior outlier removal. In these cases, the outlier removal would result in removing the model estimates that are positive and although actual ET may be quite negligible, a zero estimate is not considered to be physically realistic. However, in these scenarios, because the majority of models may predict zero, the ensemble mean will also be highly skewed towards zero making this a conservative measure to prevent zero ensemble estimates.

Statistical analyses

Key summary statistics including the least squares linear regression slope forced through the origin (slope) as well as linear regression with an intercept (Supplementary Table 7), MBE, MAE, RMSE and the coefficient of determination (r^2) were computed using paired observations between OpenET model ET estimates and post-processed and corrected flux ET estimates[19]. Daily accuracy statistics were not compared against any gap-filled station ET data, and monthly statistics only used station ET with 5 or fewer gap-filled days per month. Growing season and annual evaluations used paired monthly data and did not include any periods with monthly gaps. Also, the number of paired observations was always the same among models for all statistical analyses.

All statistics were calculated on a site-by-site basis using paired model–measured ET using the Python Numpy package version 1.17.2 (ref. 63). For linear regression, the Numpy linalg.lstsq algorithm was used, and it applies the least squares approach. We used the modelled ET as the dependent variable and the measured ET as the independent variable.

The MBE was calculated as

$$\mathrm{MBE} = \frac{1}{n}\sum_{i=1}^{n}(P_i - O_i),$$

where O_i is the observed ET, P_i is the model predicted ET and n is the total number of paired model–measured ET data points.

The MAE was calculated as

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|P_i - O_i|,$$

and the RMSE was calculated as

$$\mathrm{RMSE} = \sqrt{\sum_{i=1}^{n}\frac{(P_i - O_i)^2}{n}}.$$

Here, r^2 values were calculated as the square of the Pearson correlation coefficient, which was calculated from paired model–measurement ET data using the Python statsmodels package, version 0.12.1 (ref. 64).

For grouping statistics by land cover or climate zone we used two methods: (1) for the computation of linear regression and r^2 all data from each ground observation in a group (for example, monthly paired model–station ET estimates for annual crop stations) were pooled together before computing a single statistic per model; and (2) MBE, MAE and RMSE were computed separately for each ground station, and then a weighted mean was taken. Grouped statistics were weighted by the square root of the number of paired observations per station (n); the rationale is to avoid giving too much weight to stations with excessively long data records while also not giving equal weight to stations with short data records[65]. We also imposed data length requirements for in situ ET stations: to be included in daily grouped mean statistics we required stations to have a minimum of six paired station–model data points, and a minimum of three paired observations for inclusion in monthly grouped mean statistics. We note that Melton et al.[5] presented similar statistical metrics from a subset of cropland sites used in this study, and in that study, the linear regression slope and r^2 metrics did incorporate weighting, which we deemed inappropriate or unnecessary in this study. For congruency, the statistics computed in the same manner as in Melton et al.[5] are provided in Supplementary Table 12.

A post hoc Tukey test, also known as the honestly significant difference test, was used to compare multiple mean ET estimates from each model, the ensemble mean, and from the mean of the unclosed and closed flux ET data. The test was applied using all paired data from cropland stations, including for crop subgroups: annual crops, orchards and vineyards, at daily, monthly, growing season and annual timescales. The family-wise error rate was set to 0.05 and the test was performed using the Python statsmodels package, version 0.12.1 (ref. 64).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The in situ measured ET data analysed during the current study are available in the Zenodo repository, with identifier https://doi.org/10.5281/zenodo.7636781. The OpenET model ET data analysed during the current study are available in the Zenodo repository, with identifier https://doi.org/10.5281/zenodo.10119477.

Code availability

The code used to post-process eddy flux tower data for the current study is publicly available on GitHub (https://github.com/Open-ET/flux-data-qaqc). The code used to generate flux footprints for the current study is publicly available on GitHub (https://github.com/Open-ET/flux-data-footprint).

References

1. Fisher, J. B. et al. The future of evapotranspiration: global requirements for ecosystem functioning, carbon and climate feedbacks, agricultural management, and water resources. *Water Resour. Res.* **53**, 2618–2626 (2017).
2. Dieter, C. A. et al. Estimated use of water in the United States in 2015. *Circular* 1411 https://pubs.usgs.gov/publication/cir1441 (2018).
3. Cook, B. I., Ault, T. R. & Smerdon, J. E. Unprecedented 21st century drought risk in the American Southwest and Central Plains. *Sci. Adv.* **1**, e1400082 (2015).
4. Liu, P.-W. et al. Groundwater depletion in California's Central Valley accelerates during megadrought. *Nat. Commun.* **13**, 7825 (2022).
5. Melton, F. S. et al. OpenET: filling a critical data gap in water management for the western United States. *J. Am. Water Resour. Assoc.* **58**, 971–994 (2022).
6. Chen, J. M. & Liu, J. Evolution of evapotranspiration models using thermal and shortwave remote sensing data. *Remote Sens. Environ.* **237**, 111594 (2020).
7. Anderson, M. et al. Field-scale assessment of land and water use change over the California Delta using remote sensing. *Remote Sens.* **10**, 889 (2018).
8. Allen, R. G., Tasumi, M. & Trezza, R. Satellite-based energy balance for mapping evapotranspiration with internalized calibration (METRIC)—Model. *J. Irrig. Drain. Eng.* **133**, 380–394 (2007).
9. Laipelt, L. et al. Long-term monitoring of evapotranspiration using the SEBAL algorithm and Google Earth Engine cloud computing. *ISPRS J. Photogramm. Remote Sens.* **178**, 81–96 (2021).
10. Fisher, J. B., Tu, K. P. & Baldocchi, D. D. 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**, 901–919 (2008).
11. Pereira, L. S. et al. Prediction of crop coefficients from fraction of ground cover and height. Background and validation using ground and remote sensing data. *Agric. Water Manag.* **241**, 106197 (2020).
12. Melton, F. S. et al. Satellite irrigation management support with the terrestrial observation and prediction system: a framework for integration of satellite and surface observations to support improvements in agricultural water resource management. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **5**, 1709–1721 (2012).
13. Senay, G. B. et al. Improving the operational simplified surface energy balance evapotranspiration model using the forcing and normalizing operation. *Remote Sens.* **15**, 260 (2023).
14. Gorelick, N. et al. Google Earth Engine: planetary-scale geospatial analysis for everyone. *Remote Sens. Environ.* **202**, 18–27 (2017).
15. Allen, R. G. et al. Satellite-based energy balance for mapping evapotranspiration with internalized calibration (METRIC)—Applications. *J. Irrig. Drain. Eng.* **133**, 395–406 (2007).
16. Knipper, K. R. et al. Using high-spatiotemporal thermal satellite ET retrievals for operational water use and stress monitoring in a California vineyard. *Remote Sens.* **11**, 2124 (2019).
17. Senay, G. B., Friedrichs, M., Singh, R. K. & Velpuri, N. M. Evaluating Landsat 8 evapotranspiration for water use mapping in the Colorado River Basin. *Remote Sens. Environ.* **185**, 171–185 (2016).
18. Foster, T., Mieno, T. & Brozović, N. Satellite-based monitoring of irrigation water use: assessing measurement errors and their implications for agricultural water management policy. *Water Resour. Res.* **56**, e2020WR028378 (2020).
19. Volk, J. M. et al. Development of a benchmark eddy flux evapotranspiration dataset for evaluation of satellite-driven evapotranspiration models over the CONUS. *Agric. For. Meteorol.* **331**, 109307 (2023).
20. Volk, J. M. et al. Post-processed data and graphical tools for a CONUS-wide eddy flux evapotranspiration dataset. *Data Brief* https://doi.org/10.1016/j.dib.2023.109274 (2023).
21. Baldocchi, D. Measuring fluxes of trace gases and energy between ecosystems and the atmosphere—the state and future of the eddy covariance method. *Glob. Change Biol.* **20**, 3600–3609 (2014).
22. Baldocchi, D. et al. 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**, 2415–2434 (2001).
23. Hampel, F. R. The influence curve and its role in robust estimation. *J. Am. Stat. Assoc.* **69**, 383–393 (1974).
24. Leys, C., Ley, C., Klein, O., Bernard, P. & Licata, L. Detecting outliers: do not use standard deviation around the mean, use absolute deviation around the median. *J. Exp. Soc. Psychol.* **49**, 764–766 (2013).
25. Thompson, P. D. How to improve accuracy by combining independent forecasts. *Mon. Weather Rev.* **105**, 228–229 (1977).
26. Kirtman, B. P. et al. The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Am. Meteorol. Soc.* **95**, 585–601 (2014).
27. Bai, Y. et al. On the use of machine learning based ensemble approaches to improve evapotranspiration estimates from croplands across a wide environmental gradient. *Agric. For. Meteorol.* **298**, 108308 (2021).
28. Novick, K. A. et al. The AmeriFlux network: a coalition of the willing. *Agric. For. Meteorol.* **249**, 444–456 (2018).
29. Pastorello, G. et al. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Sci. Data* **7**, 1–27 (2020).
30. Mauder, M., Foken, T. & Cuxart, J. Surface-energy-balance closure over land: a review. *Bound. Layer Meteorol.* **177**, 395–426 (2020).
31. Ingwersen, J., Imukova, K., Högy, P. & Streck, T. On the use of the post-closure methods uncertainty band to evaluate the performance of land surface models against eddy covariance flux data. *Biogeosciences* **12**, 2311–2326 (2015).
32. Knipper, K. R. et al. Evapotranspiration estimates derived using thermal-based satellite remote sensing and data fusion for irrigation management in California vineyards. *Irrig. Sci.* **37**, 431–449 (2019).
33. Bambach, N. et al. Evapotranspiration uncertainty at micro-meteorological scales: the impact of the eddy covariance energy imbalance and correction methods. *Irrig. Sci.* **40**, 445–461 (2022).
34. Rubel, F., Brugger, K., Haslinger, K. & Auer, I. The climate of the European Alps: shift of very high resolution Köppen–Geiger climate zones 1800–2100. *Meteorol. Z.* **26**, 115–125 (2017).
35. Yang, Y. et al. Studying drought-induced forest mortality using high spatiotemporal resolution evapotranspiration data from thermal satellite imaging. *Remote Sens. Environ.* **265**, 112640 (2021).
36. Isaacson, B. N., Yang, Y., Anderson, M. C., Clark, K. L. & Grabosky, J. C. The effects of forest composition and management on evapotranspiration in the New Jersey pinelands. *Agric. For. Meteorol.* **339**, 109588 (2023).
37. Qian, Y. et al. Neglecting irrigation contributes to the simulated summertime warm-and-dry bias in the central United States. *Npj Clim. Atmos. Sci.* **3**, 31 (2020).

38. Lei, F., Crow, W. T., Holmes, T. R., Hain, C. & Anderson, M. C. Global investigation of soil moisture and latent heat flux coupling strength. *Water Resour. Res.* **54**, 8196–8215 (2018).
39. Dong, J., Lei, F. & Crow, W. T. Land transpiration–evaporation partitioning errors responsible for modeled summertime warm bias in the central United States. *Nat. Commun.* **13**, 336 (2022).
40. Abolafia-Rosenzweig, R., Pan, M., Zeng, J. & Livneh, B. Remotely sensed ensembles of the terrestrial water budget over major global river basins: an assessment of three closure techniques. *Remote Sens. Environ.* **252**, 112191 (2021).
41. Wang, Q. et al. Land surface models significantly underestimate the impact of land-use changes on global evapotranspiration. *Environ. Res. Lett.* **16**, 124047 (2021).
42. Allen, R. G., Pereira, L. S., Howell, T. A. & Jensen, M. E. Evapotranspiration information reporting: I. Factors governing measurement accuracy. *Agric. Water Manag.* **98**, 899–920 (2011).
43. Adu, M. O., Yawson, D. O., Armah, F. A., Asare, P. A. & Frimpong, K. A. Meta-analysis of crop yields of full, deficit, and partial root-zone drying irrigation. *Agric. Water Manag.* **197**, 79–90 (2018).
44. Xue, J. et al. Improving the spatiotemporal resolution of remotely sensed ET information for water management through Landsat, Sentinel-2, ECOSTRESS and VIIRS data fusion. *Irrig. Sci.* **40**, 609–634 (2022).
45. Gao, F. & Zhang, X. Mapping crop phenology in near real-time using satellite remote sensing: challenges and opportunities. *J. Remote Sens.* **2021**, 8379391 (2021).
46. Müller, M. Dynamic time warping. in *Information Retrieval for Music and Motion.* 69–84 (Springer, 2007).
47. Bambach, N. et al. The Tree-crop Remote sensing of Evapotranspiration eXperiment (T-REX): a science-based path for sustainable water management and climate mitigation. *Bull. Am. Meteorol. Soc.* In the press (2023).
48. Fisher, J. B. Hydrosat: towards daily, field-scale, global evapotranspiration from space. (2022).
49. Polhamus, A., Fisher, J. B. & Tu, K. P. What controls the error structure in evapotranspiration models? *Agric. For. Meteorol.* **169**, 12–24 (2013).
50. Blankenau, P. A., Kilic, A. & Allen, R. An evaluation of gridded weather data sets for the purpose of estimating reference evapotranspiration in the United States. *Agric. Water Manag.* **242**, 106376 (2020).
51. Doherty, C. T. et al. Effects of meteorological and land surface modeling uncertainty on errors in winegrape ET calculated with SIMS. *Irrig. Sci.* **40**, 515–530 (2022).
52. Purdy, A., Fisher, J., Goulden, M. & Famiglietti, J. Ground heat flux: an analytical review of 6 models evaluated at 88 sites and globally. *J. Geophys. Res. Biogeosci.* **121**, 3045–3059 (2016).
53. Allen, R. G. et al. A recommendation on standardized surface resistance for hourly calculation of reference ETo by the FAO56 Penman-Monteith method. *Agric. Water Manag.* **81**, 1–22 (2006).
54. Jung, M. et al. The FLUXCOM ensemble of global land–atmosphere energy fluxes. *Sci. Data* **6**, 74 (2019).
55. Reitz, M., Senay, G. B. & Sanford, W. E. Combining remote sensing and water-balance evapotranspiration estimates for the conterminous United States. *Remote Sens.* **9**, 1181 (2017).
56. Volk, J. et al. flux-data-qaqc: a Python package for energy balance closure and post-processing of eddy flux. *Data.* **6**, 1–5 (2021).
57. Evett, S. R. et al. The Bushland weighing lysimeters: a quarter century of crop ET investigations to advance sustainable irrigation. *Trans. ASABE* **59**, 163–179 (2016).
58. Abatzoglou, J. T. Development of gridded surface meteorological data for ecological applications and modelling. *Int. J. Climatol.* **33**, 121–131 (2013).
59. Kljun, N., Calanca, P., Rotach, M. W. & Schmid, H. P. A simple two-dimensional parameterisation for Flux Footprint Prediction (FFP). *Geosci. Model Dev.* **8**, 3695–3713 (2015).
60. Xia, Y. et al. Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *J. Geophys. Res. Atmos.* **117**, D03109 (2012).
61. Foga, S. et al. Cloud detection algorithm comparison and validation for operational Landsat data products. *Remote Sens. Environ.* **194**, 379–390 (2017).
62. Rousseeuw, P. J. & Croux, C. Alternatives to the median absolute deviation. *J. Am. Stat. Assoc.* **88**, 1273–1283 (1993).
63. Harris, C. R. et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
64. Seabold, S. & Perktold, J. Statsmodels: econometric and statistical modeling with Python. In *Proc. 9th Python in Science Conference* vol. 57 10–25080 (SciPy, 2010).
65. Obrecht, N. A. Sample size weighting follows a curvilinear function. *J. Exp. Psychol. Learn. Mem. Cogn.* **45**, 614 (2019).

Acknowledgements

OpenET data used in this study was produced on Google Earth Engine and we gratefully acknowledge Google, Inc. for the computing support and resources used to produce and process these data. Work on this analysis was supported by the Walton Family Foundation; Lyda Hill Philanthropies; National Aeronautics and Space Administration (NASA) Applied Science Program (grant NNX17AF53G, J.L.H.; NNX12AD05A, F.M.); United States Geological Survey (USGS) NASA Landsat Science Team (grant number 140G0118C0007, J.L.H.); USGS Cooperative Ecosystem Studies Units (CESU) (grant G23AC00568, A.P.); USGS Water Resources Research Institute (grant G22AC00584-00, J.L.H.), NASA Western Water Applications Office (grant 1669431, J.L.H.; 80NSSC23K0836, C.P.); California State University Agricultural Research Institute (grant number 21-01-106, F.M.); National Institute of Food and Agriculture McIntire Stennis (grant MISZ-721160, Y.Y.); Desert Research Institute Maki Endowment; and the Idaho Agricultural Experiment Station and Nebraska Agricultural Experiment Station. In-kind support is provided by partners in the agricultural and water management communities, Environmental Defense Fund and Google Earth Engine. We acknowledge and thank the long-term data collection efforts by the AmeriFlux program, the US Department of Agriculture (USDA) Agricultural Research Service, the USGS, and the USDA National Agricultural Statistics Service. The DOIs for the AmeriFlux and USGS stations used in this study include the following: https://doi.org/10.17190/AMF/1436327, https://doi.org/10.17190/AMF/1436328, https://doi.org/10.17190/AMF/1418680, https://doi.org/10.17190/AMF/1246137, https://doi.org/10.17190/AMF/1246025, https://doi.org/10.17190/AMF/1246026, https://doi.org/10.17190/AMF/1246027, https://doi.org/10.17190/AMF/1246028, https://doi.org/10.17190/AMF/1480317, https://doi.org/10.17190/AMF/1419513, https://doi.org/10.17190/AMF/1246040, https://doi.org/10.17190/AMF/1246031, https://doi.org/10.17190/AMF/1246032, https://doi.org/10.17190/AMF/1246036, https://doi.org/10.17190/AMF/1246038, https://doi.org/10.17190/AMF/1246039, https://doi.org/10.17190/AMF/1246043, https://doi.org/10.17190/AMF/1660339, https://doi.org/10.17190/AMF/1246156, https://doi.org/10.17190/AMF/1246117, https://doi.org/10.17190/AMF/1419512, https://doi.org/10.17190/AMF/1246045, https://doi.org/10.17190/AMF/1246046, https://doi.org/10.17190/AMF/1246047, https://doi.org/10.17190/AMF/1246119, https://doi.org/10.17190/AMF/1246050, https://doi.org/10.17190/AMF/1246053, https://doi.org/10.17190/AMF/1246054, https://doi.org/10.17190/AMF/1246051, https://doi.org/10.17190/AMF/1246052, https://doi.org/10.17190/AMF/1246056, https://doi.org/10.17190/AMF/1246057, https://doi.org/10.17190/AMF/1246058, https://doi.org/10.17190/AMF/1562389, https://doi.org/10.17190/AMF/1543379, https://doi.org/10.17190/AMF/1246065, https://doi.org/10.17190/AMF/1246066, https://doi.org/10.17190/AMF/1617696, https://doi.org/10.17190/AMF/1498745, https://doi.org/10.17190/AMF/1634882,

https://doi.org/10.17190/AMF/1246070, https://doi.org/10.17190/AMF/1660346, https://doi.org/10.17190/AMF/1246074, https://doi.org/10.17190/AMF/1246076, https://doi.org/10.17190/AMF/1246079, https://doi.org/10.17190/AMF/1246128, https://doi.org/10.17190/AMF/1617715, https://doi.org/10.17190/AMF/1617716, https://doi.org/10.17190/AMF/1246080, https://doi.org/10.17190/AMF/1246081, https://doi.org/10.17190/AMF/1246083, https://doi.org/10.17190/AMF/1419506, https://doi.org/10.17190/AMF/1480314, https://doi.org/10.17190/AMF/1246084, https://doi.org/10.17190/AMF/1246085, https://doi.org/10.17190/AMF/1246086, https://doi.org/10.17190/AMF/1246088, https://doi.org/10.17190/AMF/1246089, https://doi.org/10.17190/AMF/1246092, https://doi.org/10.17190/AMF/1418683, https://doi.org/10.17190/AMF/1246093, https://doi.org/10.17190/AMF/1419507, https://doi.org/10.17190/AMF/1419508, https://doi.org/10.17190/AMF/1419509, https://doi.org/10.17190/AMF/1617721, https://doi.org/10.17190/AMF/1617724, https://doi.org/10.17190/AMF/1375201, https://doi.org/10.17190/AMF/1419502, https://doi.org/10.17190/AMF/1419501, https://doi.org/10.17190/AMF/1419504, https://doi.org/10.17190/AMF/1246136, https://doi.org/10.17190/AMF/1246105, https://doi.org/10.17190/AMF/1246096, https://doi.org/10.17190/AMF/1418684, https://doi.org/10.17190/AMF/1246097, https://doi.org/10.17190/AMF/1246098, https://doi.org/10.17190/AMF/1246099, https://doi.org/10.17190/AMF/1246101, https://doi.org/10.17190/AMF/1246102, https://doi.org/10.17190/AMF/1246127, https://doi.org/10.17190/AMF/1246154, https://doi.org/10.17190/AMF/1246104, https://doi.org/10.17190/AMF/1418685, https://doi.org/10.17190/AMF/1660351, https://doi.org/10.17190/AMF/1246148, https://doi.org/10.17190/AMF/1246149, https://doi.org/10.17190/AMF/1246140, https://doi.org/10.17190/AMF/1245984, https://doi.org/10.17190/AMF/1246109, https://doi.org/10.17190/AMF/1246111, https://doi.org/10.17190/AMF/1246112, https://doi.org/10.17190/AMF/1617728, https://doi.org/10.17190/AMF/1579721, https://doi.org/10.17190/AMF/1617732, https://doi.org/10.17190/AMF/1579723, https://doi.org/10.17190/AMF/1617735, https://doi.org/10.17190/AMF/1617737, https://doi.org/10.17190/AMF/1617741, https://doi.org/10.3133/sir20095079, https://doi.org/10.3133/sir20095079, https://doi.org/10.3133/sir20055288, https://doi.org/10.3133/sir20055288, https://doi.org/10.3133/sir20085116, https://doi.org/10.3133/sir20095079, https://doi.org/10.3133/sir20085116, https://doi.org/10.5066/F7R49NZN, https://doi.org/10.5066/F7R49NZN, https://doi.org/10.5066/F79C6WM9, https://doi.org/10.5066/F79C6WM9, https://doi.org/10.3133/pp1805, https://doi.org/10.5066/P9NZ9XSP, https://doi.org/10.5066/P9NZ9XSP, https://doi.org/10.5066/P9NZ9XSP, https://doi.org/10.3133/sir20075078, https://doi.org/10.3133/sir20075078, https://doi.org/10.3133/sir20085116, https://doi.org/10.3133/sir20075078 and https://doi.org/10.3133/sir20075078. Funding for AmeriFlux data resources was provided by the US Department of Energy Office of Science. Any use of trade, firm or product names is for descriptive purposes only and does not imply endorsement by the US Government.

Author contributions

F.S.M., J.L.H., J.M.V., R.A., M.A., J.B.F., A.K., A.R., G.B.S. and C.P. designed and guided the study; J.M.V., F.S.M., M.A. and L.J. wrote the main text; J.M.V. performed statistical analyses; C.M., J.M.V., B.M., T.O., C.D. and T.W. prepared measured data or model input data, or ran models; F.S.M., R.A., M.A., J.B.F., A.K., A.R., G.B.S., J.L.H., C.M., W.C., C.T.D., M.F., A.G., C.H., G.H., L.J., Y.K., K.K., S.O.-S., G.E.L.P., A.P., P.R., Y.Y., L.L. and B.C.d.A. developed models and OpenET infrastructure; J.M.V., M.A., F.S.M., L.J., R.A., J.B.F., J.L.H., A.K., G.B.S., T.O., B.M., A.R., M.F. and T.W. reviewed and edited text and figures.

Competing interests

The authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s44221-023-00181-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44221-023-00181-7.

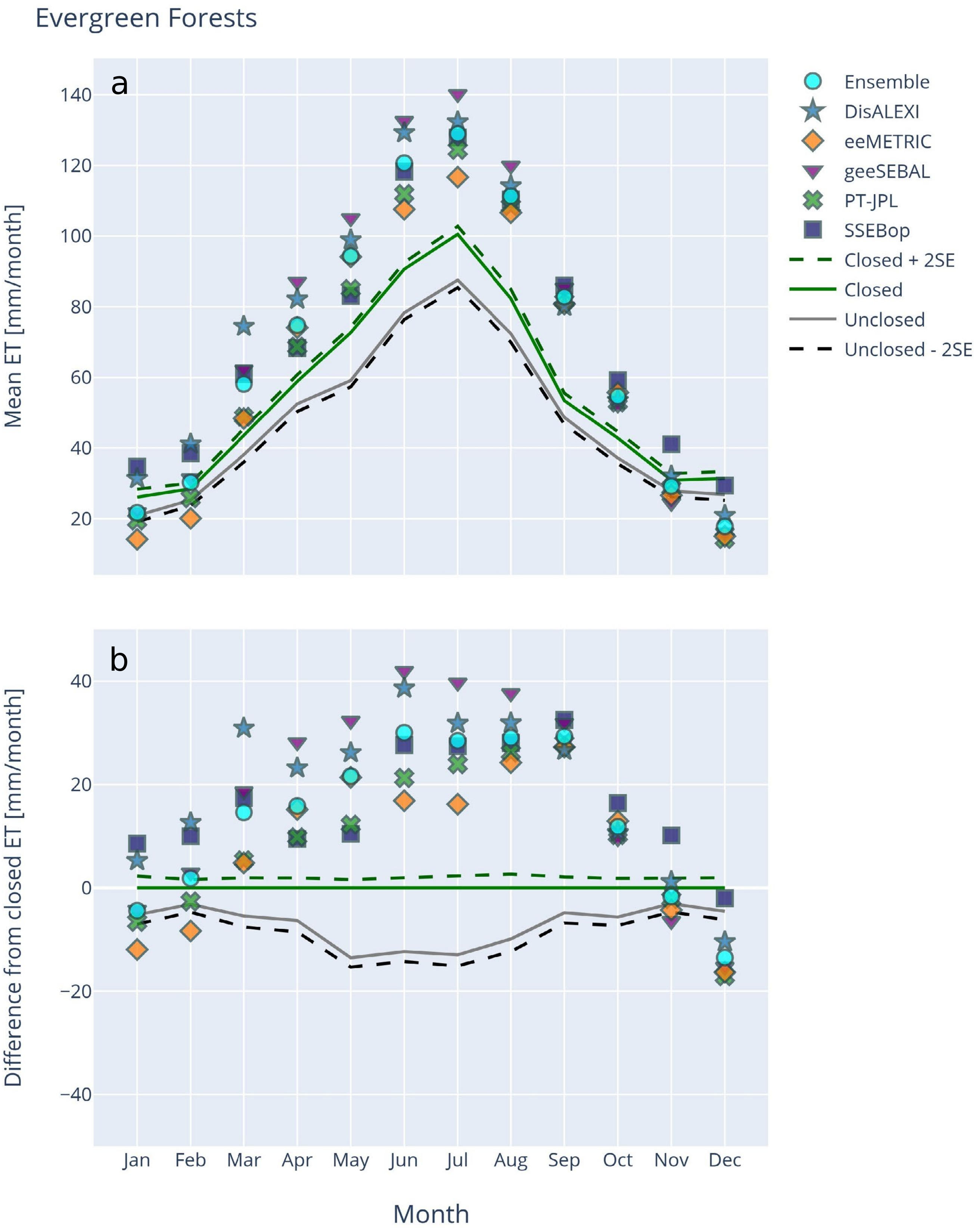
Correspondence and requests for materials should be addressed to John M. Volk.

Peer review information *Nature Water* thanks Tilden Meyers, Dennis Baldocchi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

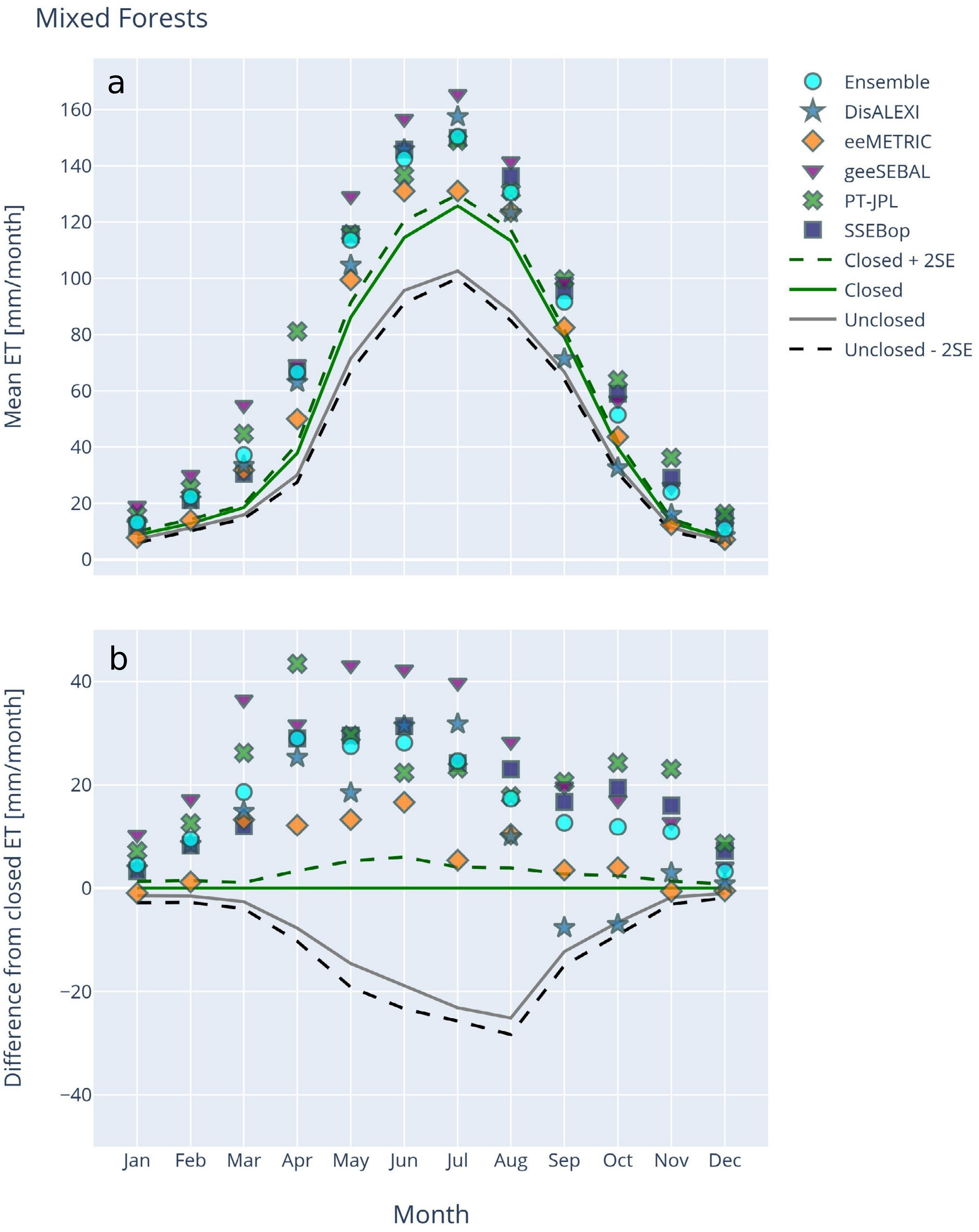
Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

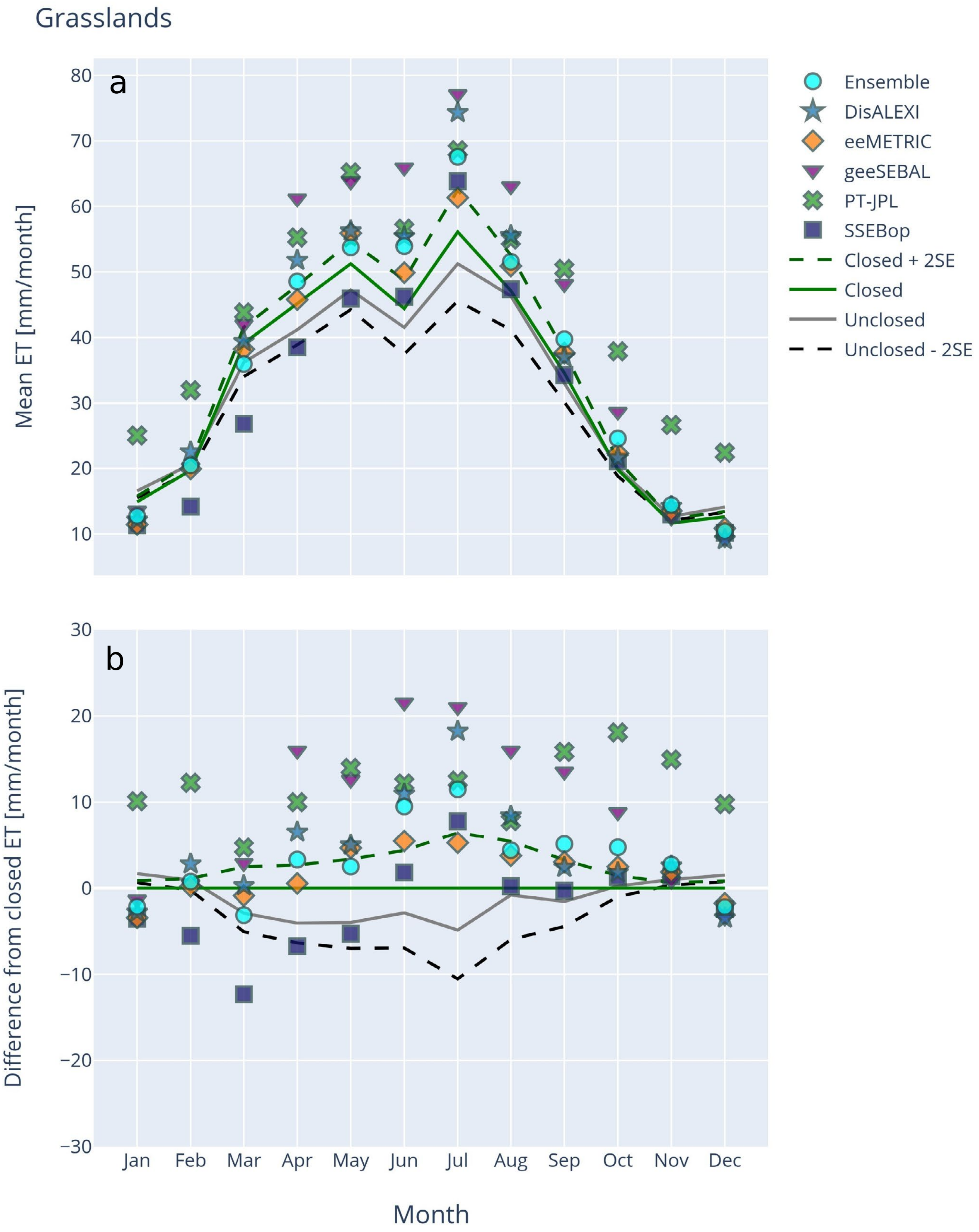
[1]Desert Research Institute, Reno, NV, USA. [2]NASA Ames Research Center, Moffett Field, CA, USA. [3]California State University Monterey Bay, Seaside, CA, USA. [4]University of Idaho, Kimberly, ID, USA. [5]USDA Agricultural Research Service, Beltsville, MD, USA. [6]University of California, Los Angeles, Los Angeles, CA, USA. [7]University of Nebraska-Lincoln, Lincoln, NE, USA. [8]Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil. [9]US Geological Survey Earth Resources Observation and Science Center, North Central Climate Adaptation Science Center, Fort Collins, CO, USA. [10]KBR, Inc. under contract to the US Geological Survey Earth Resources Observation and Science Center, Sioux Falls, SD, USA. [11]NASA Marshall Space Flight Center, Huntsville, AL, USA. [12]NASA Jet Propulsion Lab, Pasadena, CA, USA. [13]University of California Berkeley, Berkeley, USA. [14]USDA Agricultural Research Service, Davis, CA, USA. [15]Innovate!, Inc. under contract to the US Geological Survey Earth Resources Observation and Science Center, Sioux Falls, SD, USA. [16]Mississippi State University, Starkville, MS, USA. ✉e-mail: john.volk@dri.edu



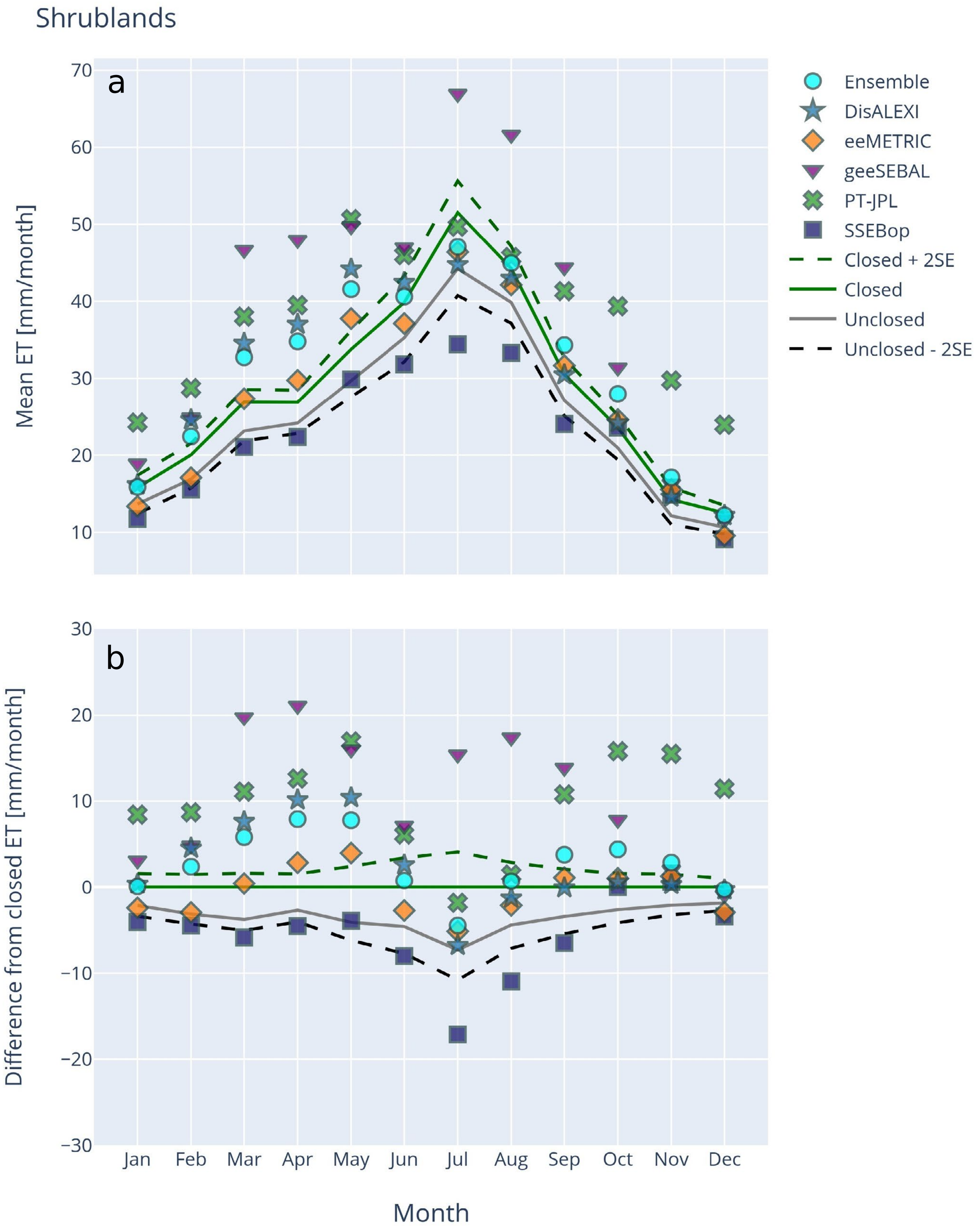
Extended Data Fig. 1 | Monthly climatology of paired modeled and observed ET for evergreen forest sites. Subplot (**a**) shows monthly climatology of paired OpenET[5] and flux tower ET[19,20] from evergreen forested sites. Subplot (**b**) shows the residual of monthly mean ET (model minus mean closed flux ET). Unclosed and closed labels refer to flux tower ET before and after energy balance closure correction. Dashed lines represent the closed flux ET mean plus two standard errors of the mean and unclosed flux ET mean minus two standard errors of the mean.



Extended Data Fig. 2 | Monthly climatology of paired modeled and observed ET for mixed forest sites. Subplot (**a**) shows monthly climatology of paired OpenET[5] and flux tower ET[19,20] from mixed forested sites. Subplot (**b**) shows the residual of monthly mean ET (model minus mean closed flux ET). Unclosed and closed labels refer to flux tower ET before and after energy balance closure correction. Dashed lines represent the closed flux ET mean plus two standard errors of the mean and unclosed flux ET mean minus two standard errors of the mean.

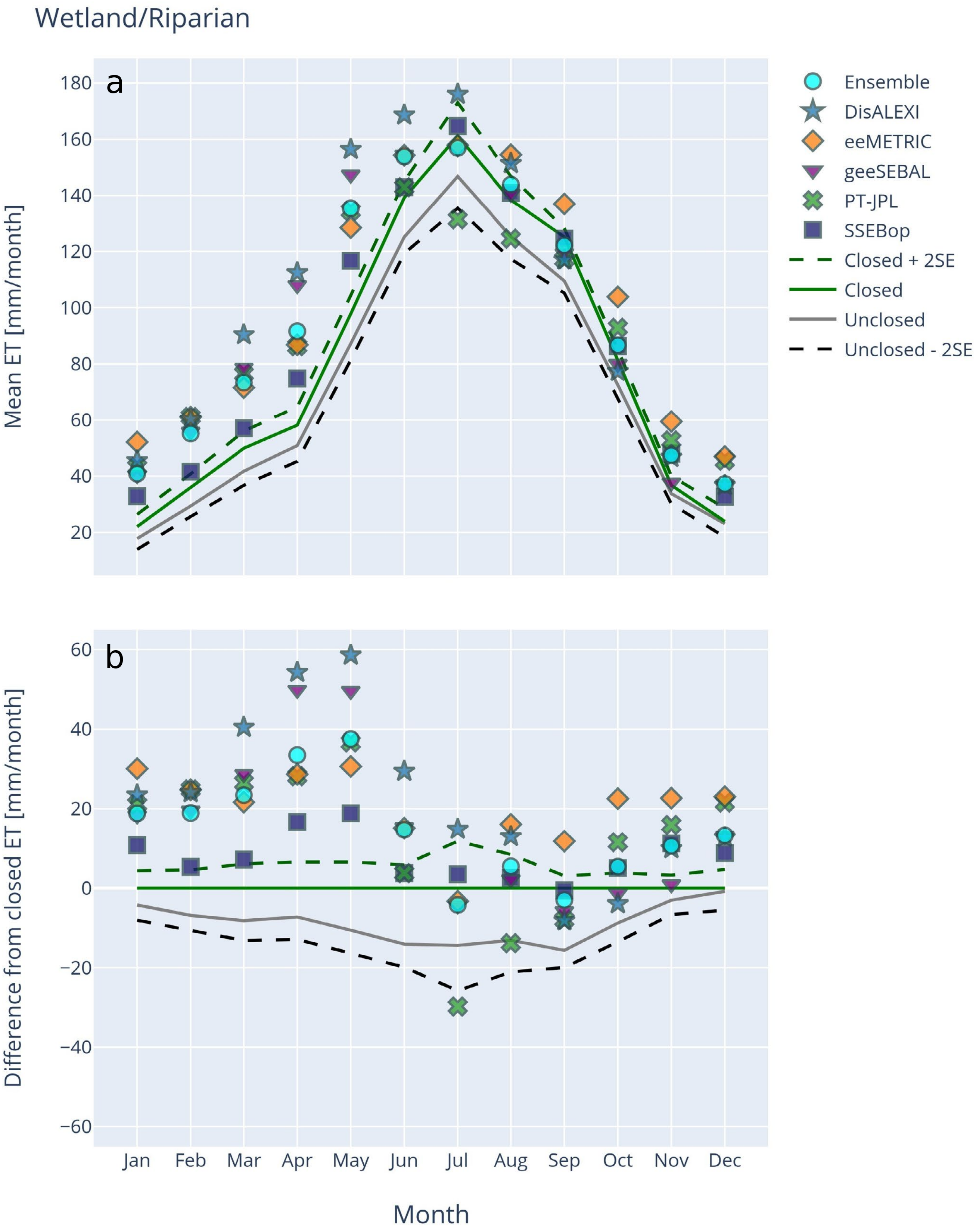


Extended Data Fig. 3 | Monthly climatology of paired modeled and observed ET for grassland sites. Subplot (**a**) shows monthly climatology of paired OpenET[5] and flux tower ET[19,20] from grassland sites. Subplot (**b**) shows the residual of monthly mean ET (model minus mean closed flux ET). Unclosed and closed labels refer to flux tower ET before and after energy balance closure correction. Dashed lines represent the closed flux ET mean plus two standard errors of the mean and unclosed flux ET mean minus two standard errors of the mean.

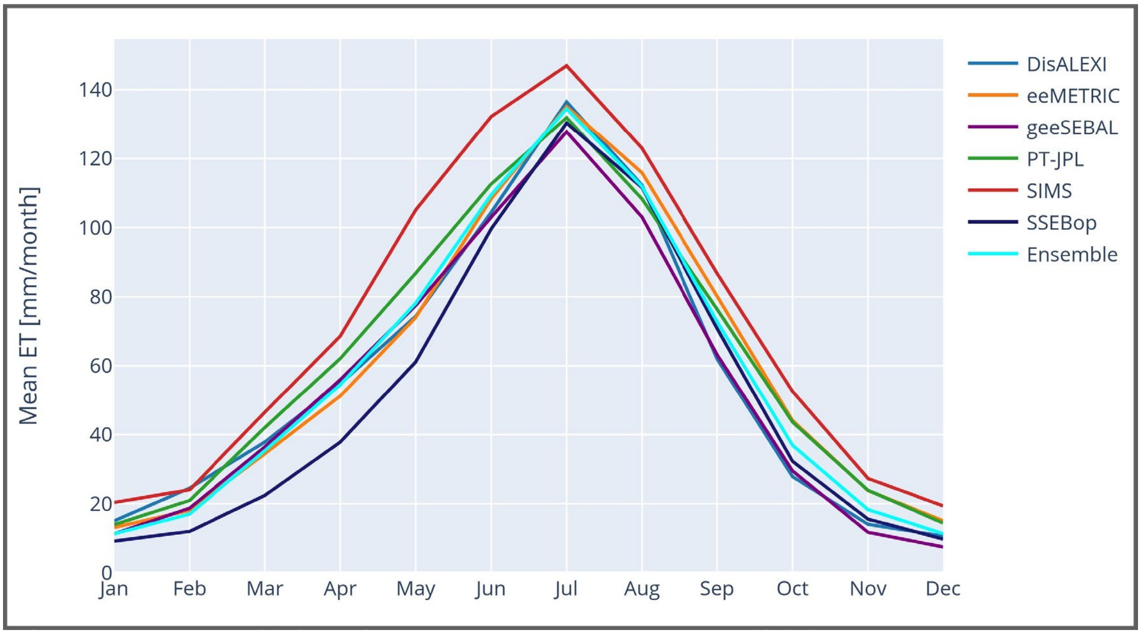


Extended Data Fig. 4 | Monthly climatology of paired modeled and observed ET for shrubland sites. Subplot (**a**) shows monthly climatology of paired OpenET[5] and flux tower ET[19,20] from shrubland sites. Subplot (**b**) shows the residual of monthly mean ET (model minus mean closed flux ET). Unclosed and closed labels refer to flux tower ET before and after energy balance closure correction. Dashed lines represent the closed flux ET mean plus two standard errors of the mean and unclosed flux ET mean minus two standard errors of the mean.

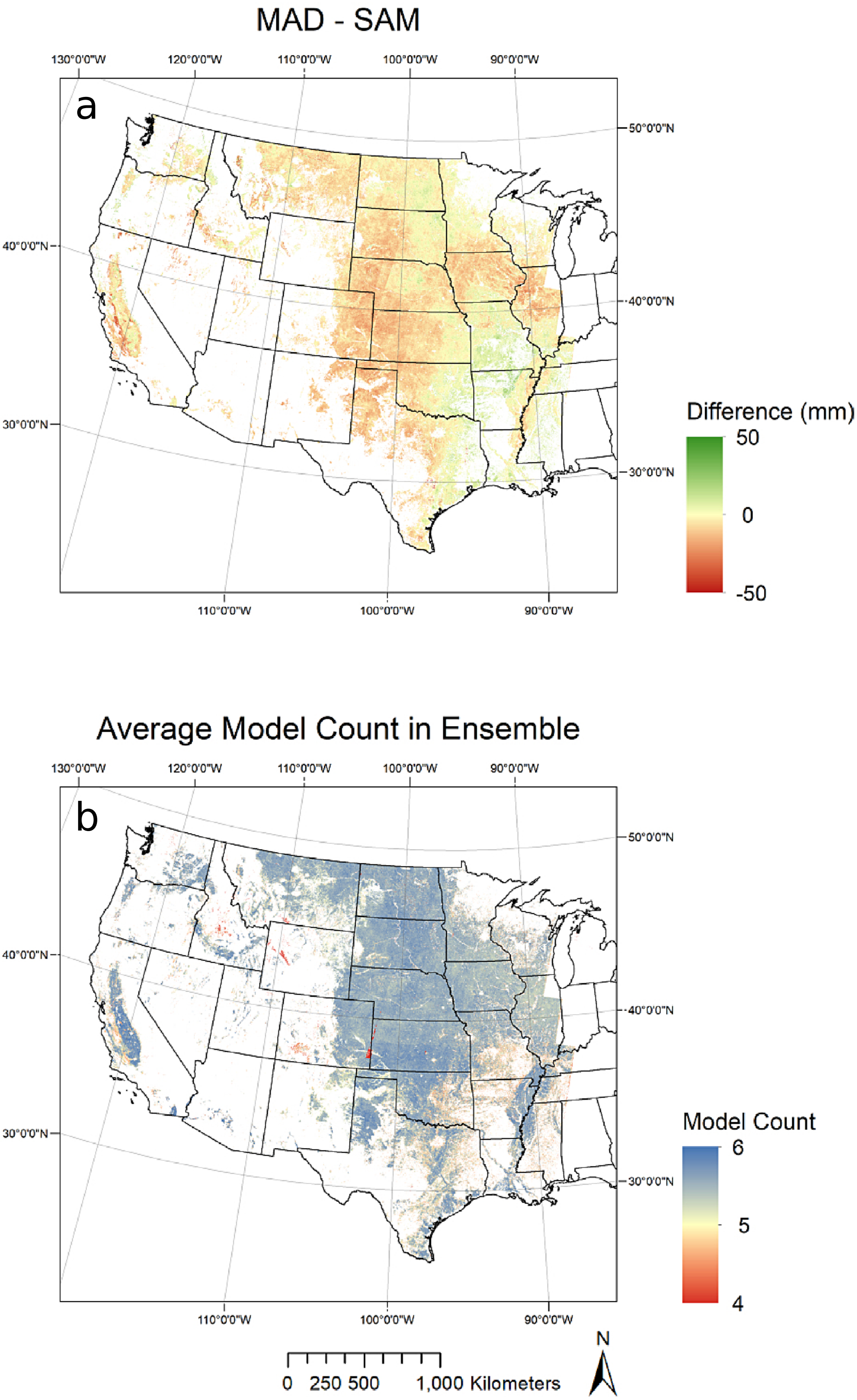
https://doi.org/10.1038/s44221-023-00181-7



Extended Data Fig. 5 | Monthly climatology of paired modeled and observed ET for wetland and riparian sites. Subplot (**a**) shows monthly climatology of paired OpenET[5] and flux tower ET[19,20] from wetland and riparian sites. Subplot (**b**) shows the residual of monthly mean ET (model minus mean closed flux ET). Unclosed and closed labels refer to flux tower ET before and after energy balance closure correction. Dashed lines represent the closed flux ET mean plus two standard errors of the mean and unclosed flux ET mean minus two standard errors of the mean.

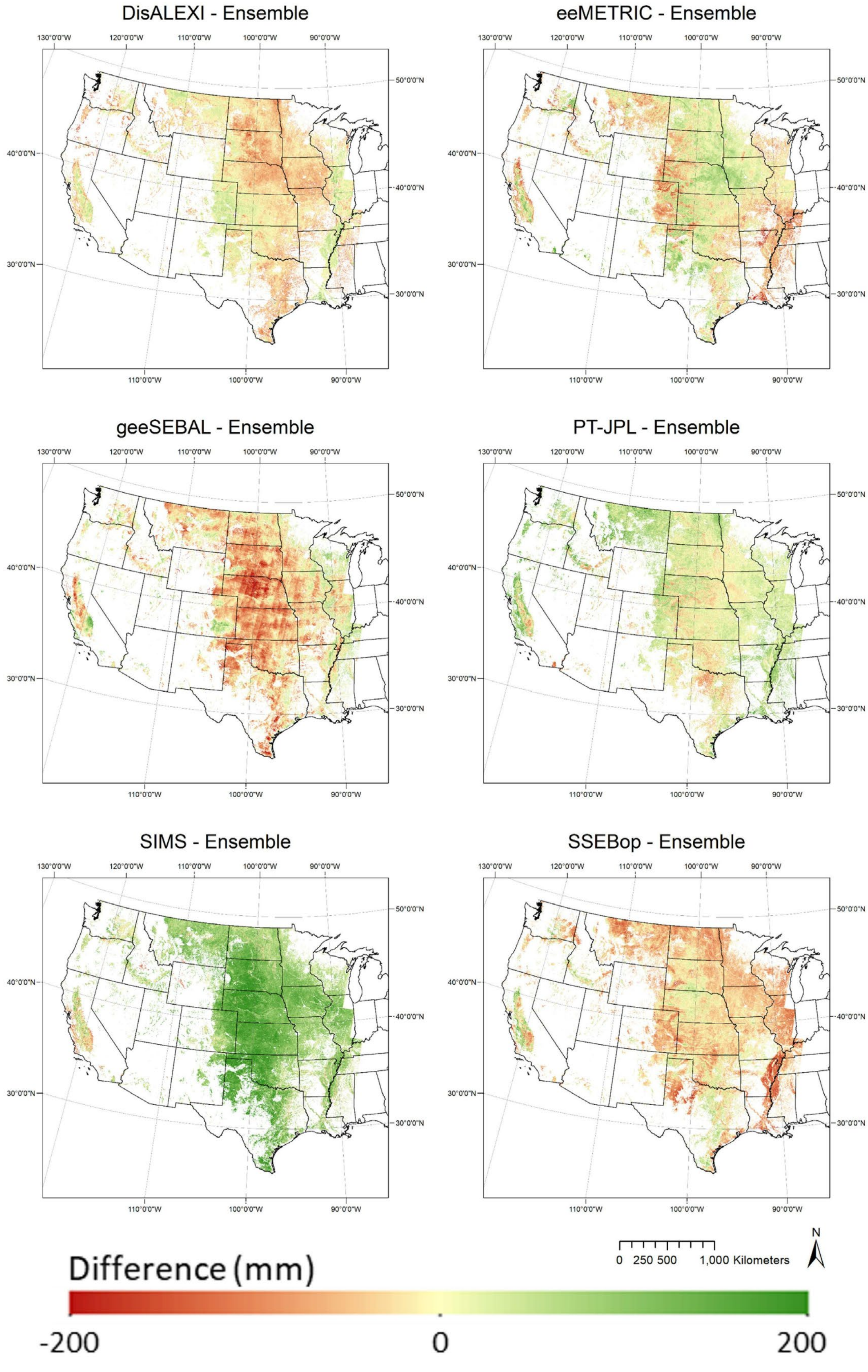


Extended Data Fig. 6 | Monthly climatology of modeled ET using all cropland pixels. Monthly climatology of OpenET[5] ensemble members and the ensemble mean using all monthly ET data for all pixels that were classified as croplands for each year from 2016–2022.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | Spatial analysis of model ensemble outlier occurrence in cropland pixels. Subplot (**a**) shows the spatial differences between the OpenET[5] ensemble mean growing season (April through October) ET for cropland pixels using the median absolute deviation (MAD) outlier removal approach and the simple arithmetic mean (SAM); monthly ET from 2016–2022 was used to build the map. Subplot (**b**) shows the average count of models used in the ensemble after outlier removal using all growing season monthly data for cropland pixels. A value of six indicates that no model was identified as an outlier, while four is the lower limit where a maximum of two models were removed as outliers before taking the ensemble mean.



Extended Data Fig. 8 | Spatial difference between mean growing season ET for each model from the ensemble value in cropland pixels. Difference between mean growing season (April through October) ET from each OpenET[5] model minus the ensemble mean using all monthly data from all pixels that were classified as croplands for each year from 2016–2022. See Supplementary Discussion 4 for a discussion of the Landsat striping exhibited by geeSEBAL.

nature portfolio

Corresponding author(s): John Volk

Last updated by author(s): Nov 27, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	□	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
□	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The in situ measured evapotranspiration data analysed during the current study were processed using the "flux-data-qaqc" Python package, version 0.1.6 (https://github.com/Open-ET/flux-data-qaqc).
Data analysis	The "flux-data-footprint" Python package was used to generate temporally dynamic flux footprints for sampling of of daily and monthly model ET data (https://github.com/Open-ET/flux-data-footprint). The Numpy (version 1.17.2) and statsmodels (version 0.12.1) Python packages were used for data analyses during the current study.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The in situ measured evapotranspiration data analysed during the current study are available in the Zenodo repository, with identifier http://dx.doi.org/10.5281/

zenodo.7636781. The OpenET model ET data analyzed during the current study are available in the Zenodo repository, with identifier http://dx.doi.org/10.5281/zenodo.10119477.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender

The current study did not involve human participants, their data, or their biological material.

Population characteristics

The current study did not involve human participants, their data, or their biological material.

Recruitment

The current study did not involve human participants, their data, or their biological material.

Ethics oversight

The current study did not involve human participants, their data, or their biological material.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences □ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description

The current study is focused on one-to-one comparisons between evapotranspiration data as modeled by remote-sensing methods and as measured on the ground; well known goodness-of-fit metrics were used to evaluate model data against measured data. Modeled and measured data were paired based on temporally overlapping records at multiple measurement stations. Accuracy metric results were grouped by different shared characteristics of the stations, using weighted averaging. Spatial long-term model data were mapped for individual models and differenced from the model ensemble average.

Research sample

The measured daily and monthly evapotranspiration data in the current study came from a public dataset that is available here: http://dx.doi.org/10.5281/zenodo.7636781. The measurements used were collected between 1995-2021. OpenET model data were generated and paired to the daily and monthly measurements, however model data was not available prior to 2001. The total number of stations with paired data was 152, and among them there were 16,444 days and 4,107 months of paired data.

Sampling strategy

As a conservative measure, we required a minimum of 3 months of paired data per measurement station to be included in weighted average accuracy metrics. In order to avoid skewing grouped average metrics, we weighted each station by the square root of the number of paired data.

Data collection

Measured ET data was previously curated and publicly archived on Zenodo (http://dx.doi.org/10.5281/zenodo.7636781). Model ET data was generated for the study using the current OpenET version.

Timing and spatial scale

Measured data was collected primarily from eddy covariance systems and post-processed to daily and monthly aggregated periods. Model ET was sampled at daily (the dates of satellite overpass) and monthly intervals at the measurement stations using pixel footprints which were determined by either long-term wind direction or from a physically based flux footprint prediction model. Flux footprints rarely exceeded the size of a 7x7 (30m resolution) grid. Gaps in measured data exist due to lapses in sensor operation or faulty data, and gaps exist in model data due to cloud coverage.

Data exclusions

No data available to us were excluded from the analyses during the current study.

Reproducibility

Data processing steps for both modeled and measured data used in the current study were made reproducible by our use of well-documented open source software. We developed Python code for eddy covariance data processing and footprint development. Similarly, the OpenET modeled data was generated using the operational models which are open source and their data are also publicly available through various mechanisms including the OpenET API and the Google Earth Engine Data Catalog. The statistical methods used in the study were limited to simple techniques, and the results were independently tested by multiple members of the OpenET group using different statistical packages such as Python and Microsoft Excel.

Randomization

Randomization of data into groups was not applicable as groups were defined by biophysical characteristics such as climate and land cover type.

Blinding

This study did not employ blinding of data due to the limited amount of high quality measured evapotranspiration data available to us. However, additional data that has not yet been compared against OpenET models was held out of this study for a future blind intercomparison and accuracy assessment of OpenET.

Did the study involve field work? □ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study	
☒	□	Antibodies
☒	□	Eukaryotic cell lines
☒	□	Palaeontology and archaeology
☒	□	Animals and other organisms
☒	□	Clinical data
☒	□	Dual use research of concern

Methods

n/a	Involved in the study	
☒	□	ChIP-seq
☒	□	Flow cytometry
☒	□	MRI-based neuroimaging