Intercomparison and Accuracy Assessment Report

OPENET

Background

A detailed intercomparison and accuracy assessment of the satellite-based ET models included in the OpenET ensemble is important so that agricultural producers and water managers understand model accuracy and limitations at field to basin scales. OpenET has conducted the largest satellite-based field-scale ET model intercomparison and accuracy assessments to date within the ET modeling community. To date, the assessment has focused on comparing satellite-based ET estimates to “in-situ” or ground-based ET estimates derived primarily from eddy covariance stations (142 stations with a minimum of 6 daily ET values and 120 stations with a minimum of three complete months of ET data). These sites are located within a variety of land uses and vegetation types across the continental U.S. and are operated and maintained by AmeriFlux, USGS, USDA, and university partners (Figure 1). In addition, the accuracy assessment included six Bowen ratio stations from shrubland sites in Nevada, and four precision weighing lysimeter datasets from cropland sites in Bushland, TX. Ongoing and future efforts will focus on intercomparisons using additional precision weighing lysimeters, publicly available state agency and grower groundwater pumping records, and watershed scale water balance estimates of ET. While this accuracy and intercomparison effort is an important step for OpenET, given that the models are being applied across a broad range of geographies and land cover types, it is worth noting that many of the models included in OpenET have been extensively assessed at local to global scales in prior studies.

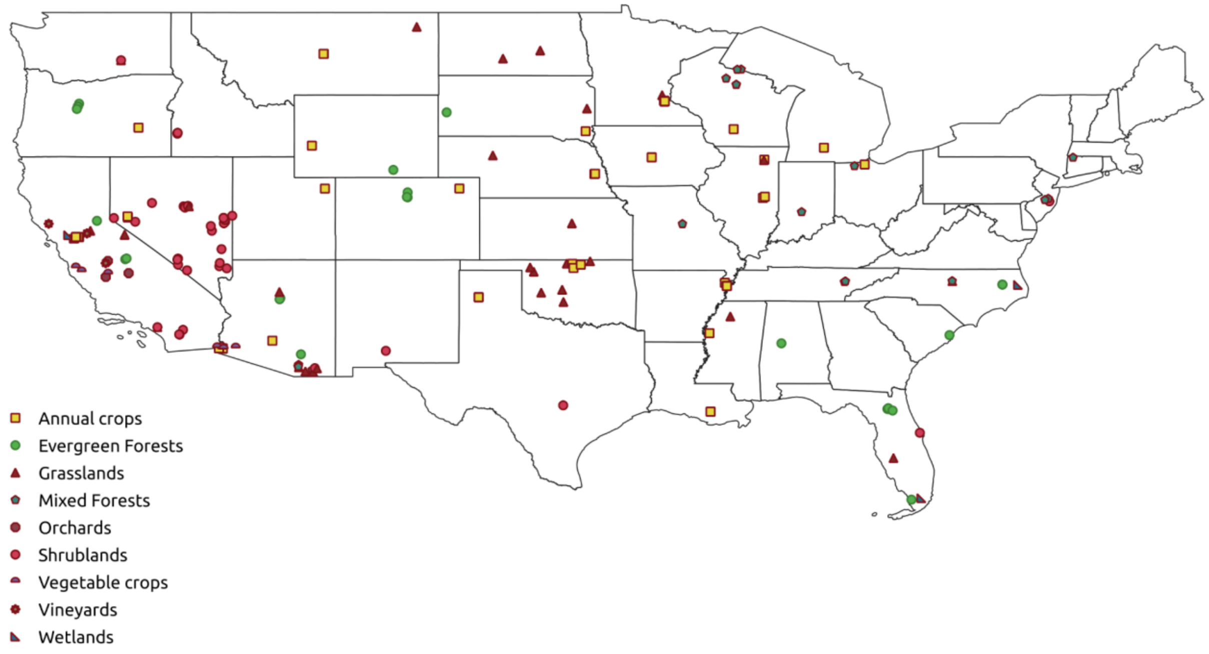

Figure 1. Map of flux tower locations included in the OpenET intercomparison and accuracy assessment.

Eddy Covariance Stations

Eddy covariance stations measure many micrometeorological variables to estimate exchanges of carbon dioxide, water vapor, and energy between the land surface and the atmosphere. Nearly all stations included in the OpenET Accuracy Assessment and Intercomparison study were instrumented with open path eddy covariance instrumentation systems (Baldocchi, 2014). While the exact instrumentation varies by site, all sites include a four-way net radiometer to measure net radiation, an infrared gas analyzer and 3D sonic anemometer to measure latent and sensible heat fluxes, and heat flux plates and soil thermocouple probes to measure ground heat fluxes. These radiation and heat fluxes are ultimately used to estimate ET. Additional information about the typical instrumentation deployed at the Ameriflux sites and estimation of ET using eddy covariance data is available in Baldocchi et al. (2001) and Foken (2008a).

Eddy covariance stations are important because they provide ground-based ET datasets that represent best available science, and that have specific locations with known spatial footprints, land use, and vegetation types. Eddy covariance station datasets were extensively assessed using a range of quality assurance and quality control and post-processing procedures. This included rigorous visual and automated screening and processing to identify outliers, fill gaps in the instrument measurement record, perform energy balance closure corrections and review of eddy covariance data following data processing procedures established by the FLUXNET organization (Pastorello et al., 2017; Pastorello et al., 2020), guidelines of Allen et al. (2011a,b), and using open source software developed by the OpenET team (Volk et al., 2021). Energy imbalance has been a significant topic of research and discussion among the micrometeorological community (Foken 2008b, Leuning et al., 2012). Energy imbalance, where available energy (i.e. net radiation minus ground heat flux) is larger than the sum of turbulent fluxes (i.e. sensible plus latent heat flux), commonly ranges between 10 to 30 percent (Foken, 2008a). There are many reasons for energy imbalance, but the most obvious is the disparity in measurement height and horizontal scales (i.e., footprint) between available energy and turbulent flux sensors. While measurement errors clearly contribute to energy imbalance they alone cannot solve the closure problem, nor can the many processing steps and corrections required to solve the closure problem. Other sources of imbalance are flux divergence due to terrain and vegetation heterogeneities, interaction of scales, and low-frequency mesoscale eddies not being captured by the instrumentation (Foken, 2008b). Energy imbalance corrections are commonly applied by researchers and practitioners even though sources of errors or the energy balance terms in question are not fully understood. Typically, turbulent fluxes are targeted through increasing latent heat flux (LE) and sensible heat flux (H) according to the Bowen ratio (Twine et al., 2000), or energy balance ratio (EBR) (Pastorello et al., 2020) as was done in this work.

Eddy covariance station location attributes were visually reviewed using aerial imagery, Landsat Normalized Difference Vegetation Index (NDVI) and wind speed and direction station data to ensure that the station location is representative of surrounding vegetation and land surface characteristics. Daily and monthly dynamic flux footprint and fetch areas for spatial sampling and intercomparison of satellite and ground-based ET datasets were developed using a two-dimensional Flux Footprint Prediction system (Kljun et al., 2015). To emphasize the flux footprint predictions during times of relatively higher evaporative demand, footprints were weighted by hourly grass reference ET using gridded weather data from NLDAS (Xia et al., 2012). Dynamic footprint areas were used to spatially sample both individual model and ensemble average ET estimates. Of the 144 eddy covariance stations included in the intercomparison, 69 stations had the necessary micrometeorological data to compute dynamic footprint areas. For most of the remaining stations, static seven by seven pixel footprints (210m by 210m) were manually selected based on wind rose diagrams (75 stations). For the remaining six stations where the selection of seven by seven pixel footprints was not feasible due to rapid changes in land cover at the edge of the footprints, static five by five footprints (150m by 150m) were selected.

In addition to daily, monthly, and annual ET data, total growing season ET was computed at each ground-based ET station and compared with satellite-based ET estimates. Growing season ET, particularly in agricultural settings, is of high importance due to the combination of high evaporative demand in the atmosphere and high water availability near the land surface. We used gridMET climate data (Abatzoglou, 2013) from 1980-2020 to define the mean annual start and end dates of the growing season at each ground-based ET station using a 300 °C cumulative degree days from January 1 to define the start date, and the first -2 °C killing frost to determine the end date similar to Huntington et al. (2016) and Allen et al. (2020). Degree days were calculated using daily average temperatures, and minimum daily temperatures were used for defining the killing frost dates using gridMET data. Monthly ET data were used to sum annual growing season totals by rounding growing season start and end dates to the nearest month, and only years without any monthly gaps for the full growing season were used.

OpenET Model and Eddy Covariance Station Intercomparison and Accuracy Assessment

The OpenET model intercomparison and accuracy assessment was conducted in two phases. For Phase I, daily and monthly ET data from each model and the ensemble average were spatially sampled over static footprints for a subset of eddy covariance stations that were randomly selected for different land cover classes, resulting in selection of 78 stations across the U.S. All models were run in a fully automated mode and the station data were not shared with the modeling teams until the comparisons were complete. For many of the models, it was the first time they had been run on a cloud-based platform in a fully automated framework over a geographic region the size of the western U.S. As such, results from Phase I were used by the modeling teams to evaluate model performance and make improvements to account for errors in the model implementation, address issues related to interoperability of gridded datasets, and address systematic errors for regions, seasons, or specific land cover types. During this time, modeling teams updated their models to use new Landsat Collection 2 at-surface reflectance and surface temperature data. Final model changes and updates were committed to respective model code repositories, and models were run to produce ET data for the Phase II intercomparison and accuracy assessment, which included all Phase I sites, with an additional 66 stations for a total of 144 flux stations plus four weighing lysimeter data records. For Phase II, daily and monthly ET data for each model and the ensemble average were spatially sampled using dynamic station footprints where they could be computed, and static station footprints for stations where micrometeorological data inputs required for computation of dynamic footprints were not available. Results from Phase II were shared with modeling teams; however, no further modifications or changes to models were made after Phase II results were shared. For Phase II intercomparison statistics, data were limited to stations with a minimum of 3 paired monthly data points and 6 paired daily data points. This requirement was imposed to limit the ability of sites with short data records to misrepresent the grouped statistics and resulted in 122 flux stations sites used in the monthly statistics and 142 flux stations in the daily. All four precision weighing lysimeter sites also met these thresholds and were included in the daily and monthly analyses. Growing season and annual ET statistical comparisons included all stations with 1 or more seasons with no monthly gaps.

Overall intercomparison and accuracy assessment results for all Phase I and Phase II sites (using dynamic and static footprints where dynamic wasn't available) are summarized in Table 1 for water year data for croplands, Table 2 for growing season data, in Table 3 for monthly data, and in Table 4 for daily data for all land covers. Results are shown for the

OpenET ensemble average with outliers dropped using the Median Absolute Deviation (MAD) approach (Hampel, 1974; Leys et al., 2013), as well as the range across the ensemble of OpenET models. Key metrics summarized in Tables 1-4 include the slope of the best fit line through the origin, the mean bias error (MBE), the mean absolute error (MAE) and the root mean squared error (RMSE) in mm/month or mm/day, and the coefficient of determination (r^2) between OpenET and closed energy balance eddy covariance station ET data. The r^2 values were calculated as the square of the Pearson's product-moment correlation coefficient. Acknowledging its limitations as a measure of goodness of fit for model evaluation (Legates and McCabe, 1999), we include r^2 as an easily interpreted indicator of the proportion of total variance in the eddy covariance station data that can be explained by the OpenET models.

To calculate the overall summary statistics in Tables 1-4 for slope, MBE, MAE and RMSE, we calculated a weighted mean value for each statistic from the ground-based ET stations using the square roots of sample size ($\sqrt{n}$) of each site following Obrecht (2019). This weighting was applied to reduce the greater influence of sites that had longer periods of record, and to ensure that model performance at all sites contributed to the overall result summaries. Since a weighted mean of r^2 values is difficult to interpret, we calculated the overall r^2 value by pooling data from all sites and then calculating the r^2 value from the pooled data.

Table 1. Overall summary of water year intercomparison and accuracy assessment metrics for Phase II results.

Land cover type	Statistic	Ensemble	Range
Croplands	Slope	0.93	0.81-1.03
14 sites	MBE (mm)	-71.61 (-7.0%)	-197.42-4.9 (-19.3-0.5%)
N = 48 water years	MAE (mm)	91.34 (8.9%)	88.91 - 199.41 (8.7 - 19.5%)
Mean station ET = 1024 (mm)	RMSE (mm)	100.48 (9.8%)	96.32 - 208.73 (9.4 - 20.4%)
	R-squared	0.98	0.93 - 0.97

The results summarized in Table 2 generally show strong overall agreement with ground-based station and lysimeter ET for the ensemble average and most models, especially for the cropland stations. The slopes of the best fit lines through the origin for croplands range from 0.88 - 1.13, with a slope of 1.00 for the ensemble average. The cropland MAE value for the ensemble average is 80.25 mm/season, which is equivalent to an average error of 13.2%. Cropland results for individual models range from 91.18 - 111.80 mm/season, which is equivalent to 15.0 - 18.4%. RMSE values for croplands, which are more strongly influenced by outlier values from each model, are 92.72 mm/season for the ensemble average and range from 108.70 - 134.31 mm/season. r^2 values for croplands show good correlation with the station ET data for all models, and range from 0.86 - 0.92 for the individual models, with a value of 0.93 for the ensemble average. These summary statistics indicate low bias errors overall, strong correlation with the eddy covariance station ET, and accuracies that are within 91.18 - 111.80 mm/season of the station ET data at the growing season timestep.

The monthly results summarized in Table 3 also show strong overall agreement with ground-based station ET for the ensemble average and most models. The slopes of the best fit lines through the origin for croplands range from 0.86 - 1.04, with a slope of 0.95 for the ensemble average. The cropland MAE value for the ensemble average is 15.55 mm/month, which is equivalent to an average error of 16.6%. Cropland results for individual models range from 17.96 - 22.92 mm/month, which is equivalent to 19.2 - 24.5%. RMSE values for croplands are 19.97 mm/month (equivalent to 0.67 mm/day) for the ensemble average and range from

Table 2. Overall summary of growing season intercomparison and accuracy assessment metrics for Phase II results.

Land cover type	Statistic	Ensemble	Range
Croplands	Slope	1	0.88 - 1.13
38 sites	MBE (mm)	-10.1 (-1.7%)	-78.61 - 47.37 (-12.9 - 7.8%)
N = 151 growing seasons	MAE (mm)	80.25 (13.2%)	91.18 - 111.8 (15.0 - 18.4%)
Mean station ET = 609 (mm)	RMSE (mm)	92.72 (15.2%)	108.7 - 134.31 (17.8 - 22.1%)
	R-squared	0.93	0.86 - 0.92
Evergreen Forests	Slope	1.24	1.0 - 1.35
14 sites	MBE (mm)	70.72 (24.7%)	-1.2 - 109.26 (-0.4 - 38.2%)
N = 87 growing seasons	MAE (mm)	92.09 (32.2%)	81.2 - 127.31 (28.4 - 44.5%)
Mean station ET = 286 (mm)	RMSE (mm)	108.91 (38.1%)	100.79 - 146.53 (35.2 - 51.2%)
	R-squared	0.89	0.82 - 0.9
Grasslands	Slope	1.18	0.86 - 1.44
19 sites	MBE (mm)	3.9 (1.8%)	-53.81 - 54.78 (-25.0 - 25.5%)
N = 79 growing seasons	MAE (mm)	83.08 (38.6%)	81.63 - 125.85 (38.0 - 58.5%)
Mean station ET = 215 (mm)	RMSE (mm)	92.05 (42.8%)	91.21 - 138.37 (42.4 - 64.4%)
	R-squared	0.8	0.53 - 0.81
Mixed Forests	Slope	1.18	0.96 - 1.34
10 sites	MBE (mm)	57.0 (21.8%)	27.8 - 82.92 (10.7 - 31.8%)
N = 38 growing seasons	MAE (mm)	60.21 (23.1%)	46.67 - 86.31 (17.9 - 33.1%)
Mean station ET = 261 (mm)	RMSE (mm)	69.73 (26.7%)	54.99 - 95.14 (21.1 - 36.5%)
	R-squared	0.97	0.92 - 0.98
Shrublands	Slope	1.06	0.71 - 1.44
21 sites	MBE (mm)	6.79 (3.9%)	-50.14 - 61.21 (-29.2 - 35.6%)
N = 75 growing seasons	MAE (mm)	70.82 (41.2%)	67.14 - 94.0 (39.0 - 54.7%)
Mean station ET = 172 (mm)	RMSE (mm)	79.21 (46.1%)	74.95 - 107.82 (43.6 - 62.7%)
	R-squared	0.56	0.3 - 0.75
Wetlands	Slope	1.18	1.02 - 1.37
7 sites	MBE (mm)	49.65 (8.3%)	5.24 - 113.37 (0.9 - 19.0%)
N = 32 growing seasons	MAE (mm)	210.18 (35.1%)	189.0 - 266.61 (31.6 - 44.6%)
Mean station ET = 598 (mm)	RMSE (mm)	262.23 (43.9%)	233.19 - 313.09 (39.0 - 52.4%)
	R-squared	0.71	0.59 - 0.75

23.43 - 28.72 mm/month (equivalent to 0.78 - 0.96 mm/day). r^2 values for croplands show strong correlation with the station ET data for all models, and range from 0.89 - 0.93 for the individual models, with a value of 0.95 for the ensemble average. These summary statistics indicate low bias errors overall, strong correlation with the eddy covariance station ET, and accuracies that are within 17.96 - 22.92 mm/month (equivalent to average of 0.60 to 0.76 mm/day) of the station ET data at a monthly timestep.

Table 3. Overall summary of monthly intercomparison and accuracy assessment metrics for Phase II results.

Land cover type	Statistic	Ensemble	Range
Croplands	Slope	0.95	0.86 - 1.04
45 sites	MBE (mm)	-3.64 (-3.9%)	-13.77 - 5.16 (-14.7 - 5.5%)
N = 1682 months	MAE (mm)	15.55 (16.6%)	17.96 - 22.92 (19.2 - 24.5%)
Mean station ET = 93.68 (mm)	RMSE (mm)	19.97 (21.3%)	23.43 - 28.72 (25.0 - 30.7%)
	R-squared	0.95	0.89 - 0.93
Evergreen Forests	Slope	1.19	1.06 - 1.32
14 sites	MBE (mm)	14.16 (23.2%)	7.07 - 21.11 (11.6 - 34.6%)
N = 783 months	MAE (mm)	23.79 (39.0%)	24.67 - 31.43 (40.4 - 51.5%)
Mean station ET = 61.02 (mm)	RMSE (mm)	28.62 (46.9%)	30.0 - 37.91 (49.2 - 62.1%)
	R-squared	0.79	0.71 - 0.78
Grasslands	Slope	1.03	0.73 - 1.28
20 sites	MBE (mm)	-1.23 (-2.9%)	-11.88 - 9.73 (-27.9 - 22.9%)
N = 672 months	MAE (mm)	19.19 (45.1%)	20.17 - 28.1 (47.4 - 66.0%)
Mean station ET = 42.56 (mm)	RMSE (mm)	24.12 (56.7%)	24.62 - 35.96 (57.8 - 84.5%)
	R-squared	0.75	0.54 - 0.8
Mixed Forests	Slope	1.16	0.99 - 1.29
10 sites	MBE (mm)	19.14 (31.4%)	8.34 - 27.06 (13.7 - 44.3%)
N = 255 months	MAE (mm)	21.54 (35.3%)	19.51 - 30.3 (32.0 - 49.7%)
Mean station ET = 61.02 (mm)	RMSE (mm)	26.8 (43.9%)	24.44 - 36.35 (40.1 - 59.6%)
	R-squared	0.88	0.8 - 0.88
Shrublands	Slope	0.95	0.65 - 1.34
24 sites	MBE (mm)	2.89 (9.3%)	-5.2 - 12.87 (-16.7 - 41.4%)
N = 681 months	MAE (mm)	15.68 (50.5%)	17.45 - 22.64 (56.2 - 72.9%)
Mean station ET = 31.07 (mm)	RMSE (mm)	19.96 (64.2%)	20.9 - 29.18 (67.3 - 93.9%)
	R-squared	0.66	0.3 - 0.69
Wetlands	Slope	1.14	1.0 - 1.25
7 sites	MBE (mm)	14.36 (16.4%)	6.71 - 25.5 (7.7 - 29.1%)
N = 269 months	MAE (mm)	28.94 (33.0%)	29.93 - 36.19 (34.2 - 41.3%)
Mean station ET = 87.57 (mm)	RMSE (mm)	35.24 (40.2%)	36.33 - 44.0 (41.5 - 50.2%)
	R-squared	0.82	0.72 - 0.81

Results for the daily data are summarized in Table 4 and are similar to the monthly results. The slopes of the best fit lines for croplands range from 0.81 - 0.94 for individual models, with a slope of 0.88 for the ensemble average. MAE values range from 0.91 - 1.14 mm/day for individual models, with a value of 0.83 for the ensemble average. RMSE values for croplands range from 1.21 - 1.46 mm/day for individual models, with a value of 1.08 mm/day for the ensemble mean. r^2 values also show good correlation with the station ET for all models, and range from 0.68 - 0.77 for individual models, with a value of 0.81 for the ensemble mean. As expected with linear interpolation of the fraction of reference ET between image dates, MAE and RMSE values increase slightly at a daily timestep, the r^2 values decrease slightly, and the slopes of the best-fit lines move away from the 1:1 line. However, taken together, these

Table 4. Overall summary of daily intercomparison and accuracy assessment metrics for Phase II results.

Land cover type	Statistic	Ensemble	Range
Croplands	Slope	0.88	0.81 - 0.94
49 sites	MBE (mm)	-0.27 (-7.4%)	-0.61 - 0.04 (-16.8 - 1.1%)
N = 4913 days	MAE (mm)	0.83 (22.8%)	0.91 - 1.14 (25.0 - 31.3%)
Mean station ET = 3.64 (mm)	RMSE (mm)	1.08 (29.7%)	1.21 - 1.46 (33.2 - 40.1%)
	R-squared	0.81	0.68 - 0.77
Evergreen Forests	Slope	1.16	0.98 - 1.29
17 sites	MBE (mm)	0.64 (27.4%)	0.17 - 0.91 (7.3 - 38.9%)
N = 1757 days	MAE (mm)	1.0 (42.7%)	1.02 - 1.35 (43.6 - 57.7%)
Mean station ET = 2.34 (mm)	RMSE (mm)	1.24 (53.0%)	1.21 - 1.64 (51.7 - 70.1%)
	R-squared	0.55	0.41 - 0.52
Grasslands	Slope	0.9	0.72 - 1.09
28 sites	MBE (mm)	-0.11 (-6.0%)	-0.38 - 0.28 (-20.7 - 15.2%)
N = 3938 days	MAE (mm)	0.83 (45.1%)	0.8 - 1.24 (43.5 - 67.4%)
Mean station ET = 1.84 (mm)	RMSE (mm)	1.08 (58.7%)	1.0 - 1.62 (54.3 - 88.0%)
	R-squared	0.54	0.21 - 0.58
Mixed Forests	Slope	1.07	0.88 - 1.19
14 sites	MBE (mm)	0.59 (26.0%)	0.01 - 0.9 (0.4 - 39.6%)
N = 1241 days	MAE (mm)	0.9 (39.6%)	0.87 - 1.28 (38.3 - 56.4%)
Mean station ET = 2.27 (mm)	RMSE (mm)	1.16 (51.1%)	1.16 - 1.62 (51.1 - 71.4%)
	R-squared	0.75	0.54 - 0.76
Shrublands	Slope	0.8	0.59 - 1.1
26 sites	MBE (mm)	0.01 (0.9%)	-0.24 - 0.36 (-20.9 - 31.3%)
N = 3223 days	MAE (mm)	0.64 (55.7%)	0.67 - 1.01 (58.3 - 87.8%)
Mean station ET = 1.15 (mm)	RMSE (mm)	0.84 (73.0%)	0.83 - 1.32 (72.2 - 114.8%)
	R-squared	0.49	0.31 - 0.49
Wetlands	Slope	1.07	0.99 - 1.16
8 sites	MBE (mm)	0.42 (13.2%)	0.15 - 0.81 (4.7 - 25.6%)
N = 931 days	MAE (mm)	1.1 (34.7%)	1.18 - 1.34 (37.2 - 42.3%)
Mean station ET = 3.17 (mm)	RMSE (mm)	1.34 (42.3%)	1.45 - 1.68 (45.7 - 53.0%)
	R-squared	0.71	0.54 - 0.64

summary statistics indicate relatively low bias errors overall, and strong correlation with station ET at daily, monthly, and seasonal timesteps.

Overall, the OpenET ensemble average performs as well or better than any individual model across nearly all metrics, and generally has the lowest MAE and RMSE values and the highest r^2 across all land cover classes, and at daily, monthly, and seasonal timesteps. The MAE for the mean of the model ensemble as a percent of the monthly n-weighted station ET ranges between 50.5% and 16.6%, with croplands being 16.6% at a monthly timestep. The MAE for daily timesteps ranges between 22.8% and 55.7% with croplands being 22.8%. For reference, members of the OpenET user working groups specified an error of ± 10-20% as the accuracy target for ET data at a monthly timestep, and ± 15-25% as the accuracy target for daily ET data.

Strength of the Ensemble Approach, and Next Steps

It is noteworthy that MAE and RMSE values for the OpenET ensemble average are lowest, or at the low end of the range of values from the individual models. One reason for the strong overall performance of the model ensemble is that individual models may occasionally "miss", and provide estimates that differ substantially from the station ET or other reference dataset. This can be due to data quality issues in input data or physical conditions that depart from the model assumptions. However, since the ensemble value is currently calculated as the average of all or a subset of models with outliers removed according to the MAD approach, and due to very different designs of the models, errors from any one model are dampened in the ensemble average, resulting in fewer large "misses" and lower MAE and RMSE values for the ensemble average.

In considering these results, however, it is important to note that most cropland sites were located in expansive regions with well-watered crops. In these regions, (including most of California's Central Valley and Delta, and most agricultural regions in the Midwest), the ensemble value appears to provide the most reliable and stable estimate of ET. However, when looking at the limited number of cropland flux stations located in arid environments, there is evidence that some models have a relatively consistent low bias for smaller agricultural areas surrounded by dry lands. In these areas, the MAD outlier filtering approach does not filter outliers as desired due to the large range in the ensemble ET values relative to the ensemble median, resulting in a low bias in the ensemble average ET value. Over the coming months, the team will continue to make improvements to the ensemble of models in these more challenging settings.

The results from the OpenET intercomparison and accuracy assessment highlight the value of using an ensemble of models to detect and remove outliers and facilitate calculation of a single ensemble value that, in many cases, has a higher accuracy than any individual model within the ensemble. They also highlight the ability to easily compare model results at scale, which has increased transparency, accelerated the ability of the ET modeling community to identify and understand differences across the ensemble, and to identify and minimize errors and biases in the ensemble as it evolves over time.

Participation of a sizable community of scientists working collaboratively has been essential to the success and lessons learned in this first intercomparison effort, and will continue to be important for rapidly advancing the science and improving the ensemble across all settings and land cover types over the coming year. As the ensemble calculation evolves, additional comparisons will be conducted and this report will be updated, along with our Best Practices Manual.

References

Abatzoglou, J.T. 2013. Development of gridded surface meteorological data for ecological applications and modelling. International Journal of Climatology, 33(1), 121-131.

Allen, R.G., Pereira, L.S., Howell, T.A. and Jensen, M.E., 2011a. Evapotranspiration information reporting: I. Factors governing measurement accuracy. Agricultural Water Management, 98(6), pp.899-920.

Allen, R.G., Pereira, L.S., Howell, T.A. and Jensen, M.E., 2011b. Evapotranspiration information reporting: II. Recommended documentation. Agricultural Water Management, 98(6), pp.921-929.

Allen, R.G., Robison, C.W., Huntington, J., Wright, J.L. and Kilic, A., 2020. Applying the FAO-56 Dual Kc Method for Irrigation Water Requirements over Large Areas of the Western US. Transactions of the ASABE, 63(6), pp.2059-2081.

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

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

Foken, T., 2008a. Micrometeorology (Vol. 308). Berlin: Springer.

Foken, T., 2008b. The energy balance closure problem: an overview, Ecological Applications, 18: 1351–1367, 2008b, doi:10.1890/06-0922.1.

Hampel, F.R., 1974. The influence curve and its role in robust estimation. Journal of the American Statistical Association, 69(346), pp.383-393.

Huber, P.J. and Ronchetti, E.M., 1981. Robust statistics. John Wiley & Sons. New York, 1(1).

Huntington, J., Morton, C., McEvoy, D., Bromley, M., Hedgewisch, K., Allen, R., and S. Gangopadhyay. 2016. Historical and Future Irrigation Water Requirements for Selected Reclamation Project Areas, Western U.S. Desert Research Institute Publication, DOI: 10.13140/RG.2.2.23078.93761, 87 p.

Kljun, N., Calanca, P., Rotach, M.W. and Schmid, H.P., 2015. A simple two-dimensional parameterisation for Flux Footprint Prediction (FFP). Geoscientific Model Development, 8(11), pp.3695-3713.

Leys, C., Ley, C., Klein, O., Bernard, P. and Licata, L., 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology, 49(4), pp.764-766.

Legates, D.R. and McCabe Jr, G.J., 1999. Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. Water Resources Research, 35(1), pp.233-241.

Leuning, R., E. Van Gorsel, W. J. Massman, and P. R. Isaac, 2012, Reflections on the surface energy imbalance problem, Agricultural and Forest Meteorology, 156: 65–74, 2012, doi:10.1016/j.agrformet.2011.12.002.

Obrecht, N.A., 2019. Sample size weighting follows a curvilinear function. Journal of Experimental Psychology: Learning, Memory, and Cognition, 45(4), p.614.

Pastorello, G., Papale, D., Chu, H., Trotta, C., Agarwal, D., Canfora, E., Baldocchi, D. and Torn, M., 2017. The FLUXNET2015 dataset: The longest record of global carbon, water, and energy fluxes is updated. Eos, 98(10.1029).

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M. and Isaac, P., 2020. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. Scientific data, 7(1), pp.1-27.

Twine, T., Kustas, W.P., Norman, J., Cook, D., Houser, P., Teyers, T.P., Prueger, J.H., Starks, P., and Wesely, M. 2000. Correcting Eddy-Covariance Flux Underestimates over a Grassland. Agricultural and Forest Meteorology, 103(3), pp.279-300.

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

Volk, J., Huntington, J., Allen, R., Melton, F., Anderson, M. and Kilic, A., 2021. flux-data-qaqc: A Python Package for Energy Balance Closure and Post-Processing of Eddy Flux Data. Journal of Open Source Software, 6(66), p.3418.