Supplementary Information for

Toward Sustainable Groundwater Management: Harnessing Remote Sensing and Climate Data to Estimate Field-Scale Groundwater Pumping and Irrigation Efficiencies

Thomas J. Ott [a*], Sayantan Majumdar [a*], Justin L. Huntington [a†], Christopher Pearson [a], Matt Bromley [a], Blake A. Minor [a], Charles G. Morton [a], Sachiko Sueki [b], Jordan P. Beamer [c], Richard Jasoni [a]

[a] Desert Research Institute, Reno, NV, USA
[b] Desert Research Institute, Las Vegas, NV, USA
[c] Oregon Water Resources Department, Salem, OR, USA

This supplementary information file has five figures and two tables referenced in the main manuscript.

Supplementary Table 1. Ensemble machine learning (ML) models and the hyperparameters tuned in a randomized grid search with five-fold cross-validation. The random seed value is set to 1234 throughout, and the root mean square error (RMSE) is used as the objective function across these models. ERT and RF are available from scikit-learn, and GBT is available from LightGBM.

Model	Hyperparameter values	Tuned hyperparameters
Extremely Randomized Trees (ERT)	'n_estimators': [300, 400, 500, 800] 'max_features': [5, 6, 7, 10, 12, 20, 30, None] 'max_depth': [8, 15, 20, 6, 10, None] 'min_samples_leaf': [1, 2] 'max_samples': [None, 0.9] 'max_leaf_nodes': [16, 20, 31, 32, 63, 127, 15, 255, 7, None] 'min_samples_split': [2, 3, 4, 0.01] *Fixed parameters*: bootstrap=True	'n_estimators': 800 'max_features': None 'max_depth': 10 'min_samples_leaf': 2 'max_samples': None 'max_leaf_nodes': 127 'min_samples_split': 2
Gradient Boosting Machine (GBT)	'n_estimators': [300, 400, 500, 800] 'max_depth': [8, 15, 20, 6, 10, -1] 'learning_rate': [0.01, 0.005, 0.05, 0.1] 'subsample': [1, 0.9, 0.8] 'colsample_bytree': [1, 0.9] 'colsample_bynode': [1, 0.9] 'path_smooth': [0, 0.1, 0.2] 'num_leaves': [16, 20, 31, 32, 63, 127, 15, 255, 7] 'min_child_samples': [30, 40, 10, 20] *Fixed parameters*: tree_learner='feature', deterministic=True, force_row_wise=True	'n_estimators': 800 'max_depth': 8 'learning_rate': 0.01 'subsample': 0.8 'colsample_bytree': 0.9 'colsample_bynode': 1 'path_smooth': 0.2 'num_leaves': 7 'min_child_samples': 20
Random Forests (RF)	Same as ERT	'n_estimators': 500 'max_features': 20 'max_depth': 6 'min_samples_leaf': 2 'max_samples': None 'max_leaf_nodes': 16 'min_samples_split': 4

* These two authors have contributed equally.
† Corresponding author at: Division of Hydrologic Sciences, Desert Research Institute, 2215 Raggio Parkway, Reno, Nevada 89512-1095, USA
Email address: justin.huntington@dri.edu (J.L. Huntington)

Supplementary Table 2. Description of the 28 predictors used in the full machine learning models to estimate groundwater pumping depths in Diamond Valley, Nevada[‡]. The data references are in the main manuscript.

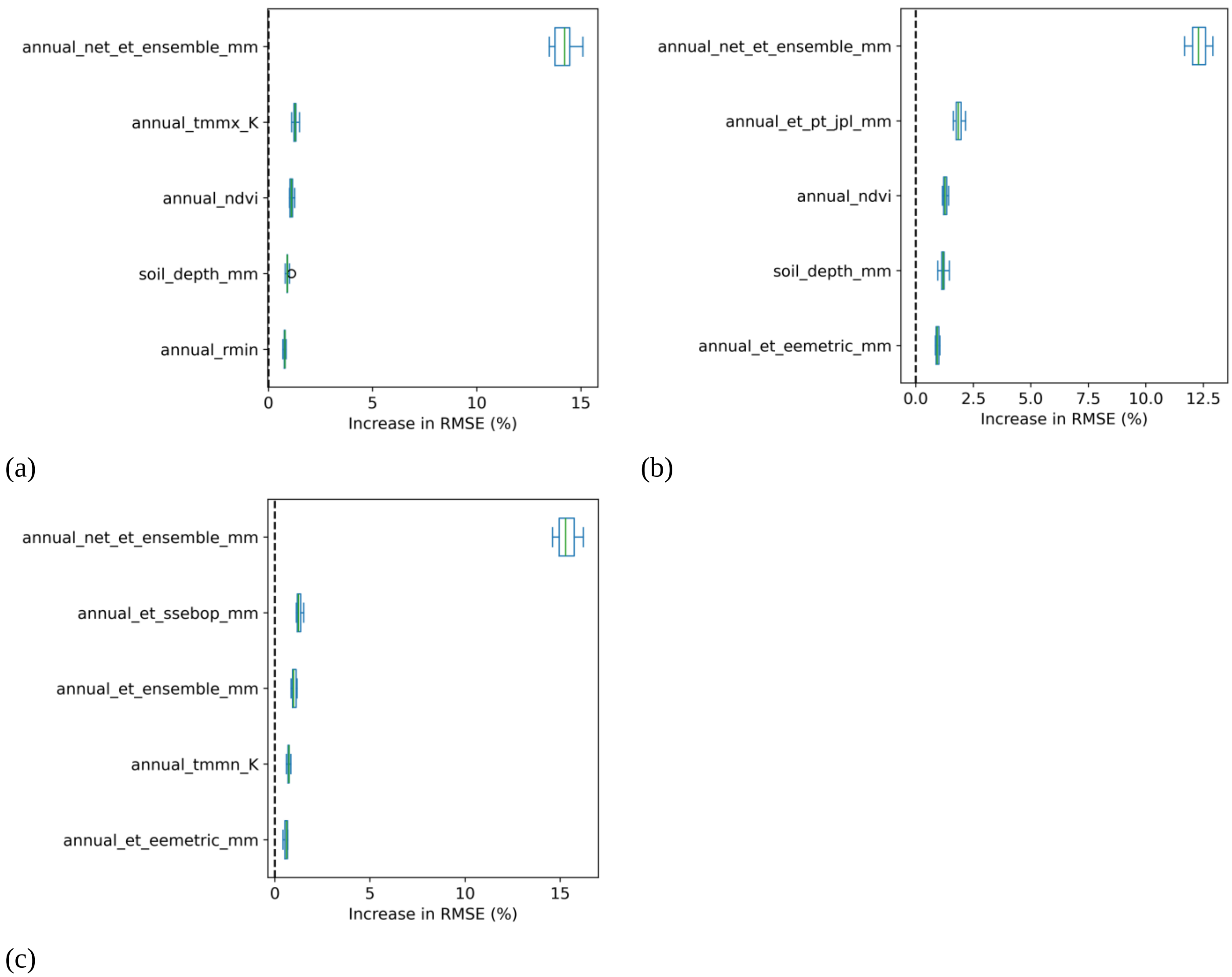
Predictor name	Description	Operations
annual_net_et_ensemble_mm[§]	OpenET ensemble-based $Net\ ET$ in mm	annual_et_ensemble_mm - annual_gridmet_precip_eff_mm
annual_et_eemetric_mm	eeMETRIC actual ET in mm	Temporal sum (calendar year) and zonal mean
annual_et_ssebop_mm	SSEBop actual ET in mm	Temporal sum (calendar year) and zonal mean
annual_et_geesebal_mm	geeSEBAL actual ET in mm	Temporal sum (calendar year) and zonal mean
annual_et_ensemble_mm	OpenET ensemble actual ET in mm	Temporal sum (calendar year) and zonal mean
annual_daymet_precip_eff_mm	Daymet v4 effective precipitation in mm	annual_daymet_precip_mm * eff_factor
annual_daymet_precip_mm	Daymet v4 precipitation in mm	Temporal sum (calendar year) and zonal mean
annual_et_disalexi_mm	ALEXI/DisALEXI actual ET in mm	Temporal sum (calendar year) and zonal mean
annual_gridmet_precip_mm	gridMET precipitation in mm	Temporal sum (calendar year) and zonal mean
annual_gridmet_precip_eff_mm	gridMET effective precipitation in mm	annual_gridmet_precip_mm * eff_factor
annual_et_pt_jpl_mm	PT-JPL actual ET in mm	Temporal sum (calendar year) and zonal mean
annual_et_sims_mm	SIMS actual ET in mm	Temporal sum (calendar year) and zonal mean
annual_ndvi	Landsat-8 32-day composite NDVI	Temporal max (calendar year) and zonal mean
annual_rmin	gridMET minimum relative humidity %	Temporal median (calendar year) and zonal mean

[‡] *Here, the temporal operations are performed for each year between 2018 and 2022, and the zonal operations are performed for each field. If a well waters multiple fields, then we sum up the corresponding actual ET, reference ET, Net ET, precipitation, effective precipitation, effective precipitation factor, and vapor pressure deficit for those fields, average the NDVI, minimum relative humidity, maximum relative humidity, minimum air temperature, maximum air temperature, soil depth, saturated hydraulic conductivity, and wind velocity, and take the mode of the hydrologic soil groups for those fields.*

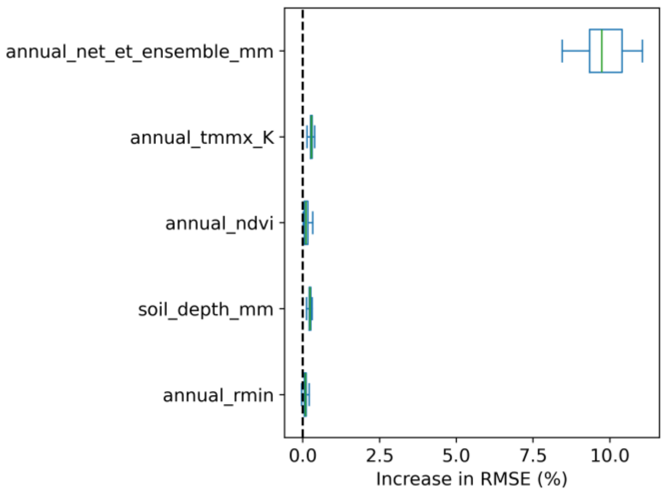
[§] *For the ML models used to compare the ET model performances (Table 5, main manuscript), we replace the annual_net_et_ensemble_mm with the corresponding Net ET (e.g., annual_net_et_eemetric_mm) and only keep the corresponding actual ET predictor, e.g, annual_et_eemetric_mm. Other ET predictors are removed to negate the correlation effects. All the remaining predictors are kept as in the full ML model. Therefore, we end up with 22 predictors for each of the models in Table 5 of the main manuscript.*

Supplementary Table 2 (Contd.). Description of the 28 predictors used in the full machine learning models to estimate groundwater pumping depths in Diamond Valley, Nevada. Here, the temporal operations are performed for each year between 2018 and 2022, and the zonal operations are performed for each field.

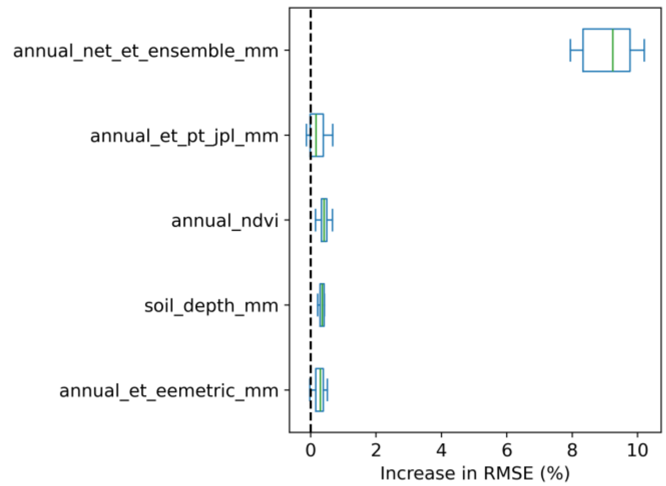
Predictor name	Description	Operations
annual_rmax	gridMET maximum relative humidity %	Temporal median (calendar year) and zonal mean
ksat_mean_micromps	Saturated hydraulic conductivity in μm/s	Zonal mean
soil_depth_mm	Soil depth in mm	Zonal mean
annual_vpd_kPa	gridMET vapor pressure deficit in kPa	Temporal sum (calendar year) and zonal mean
annual_tmmn_K	gridMET minimum air temperature (K)	Temporal median (calendar year) and zonal mean
annual_tmmx_K	gridMET maximum air temperature (K)	Temporal median (calendar year) and zonal mean
eff_factor	ET-Demands-derived basin-scale effective precipitation factor	
elevation_m	NASADEM elevation in m	Zonal mean
annual_vs_mps	gridMET wind velocity in m/s	Temporal mean (calendar year) and zonal mean
annual_etr_mm	gridMET alfalfa reference ET in mm	Temporal sum (calendar year) and zonal mean
annual_eto_mm	gridMET grass reference ET in mm	Temporal sum (calendar year) and zonal mean
HSG_1	Hydrologic soil group 1 (A)	Zonal mode
HSG_3	Hydrologic soil group 3 (B)	Zonal mode
HSG_5	Hydrologic soil group 5 (C)	Zonal mode



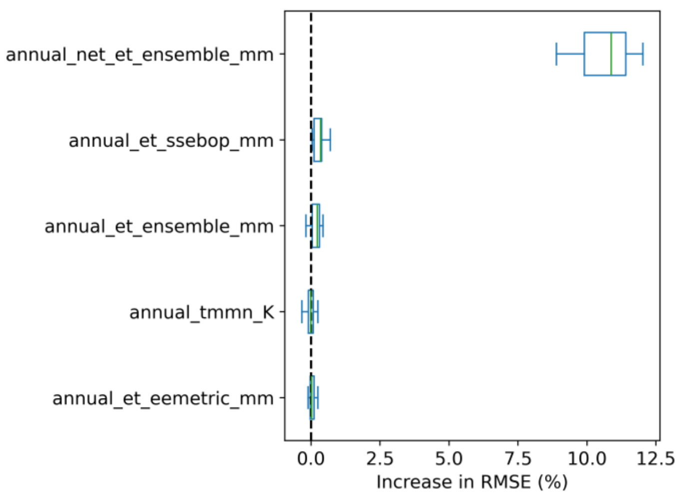
Supplementary Figure 1. Permutation importance plots showing the top five features for the training data (including validation) for (a) ERT, (b) GBM, and (c) RF.



(a)

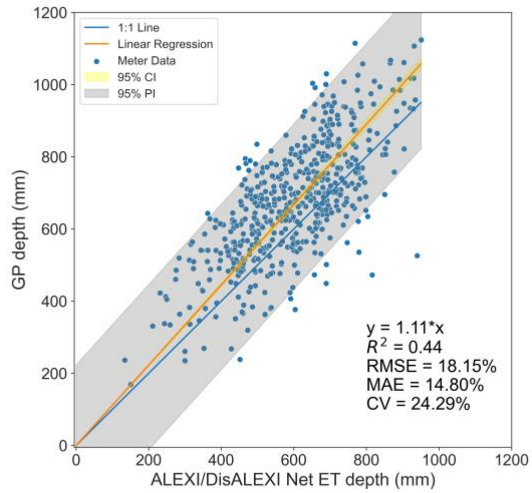


(b)

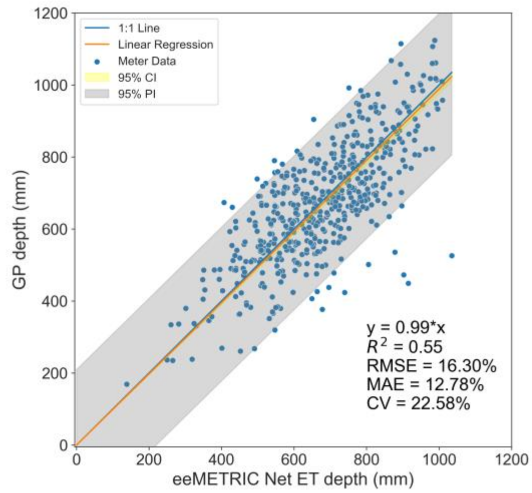


(c)

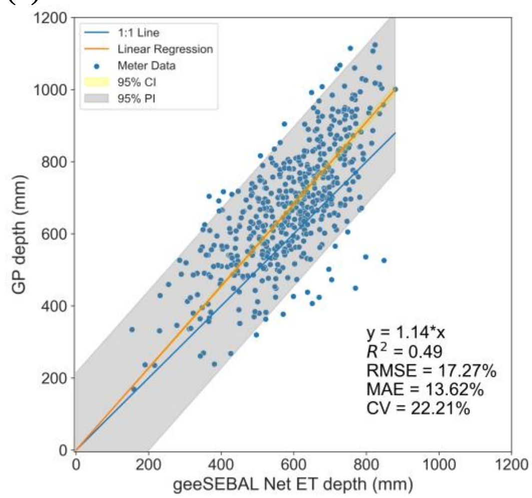
Supplementary Figure 2. Permutation importance plots showing the top five features for the test data for (a) ERT, (b) GBM, and (c) RF.



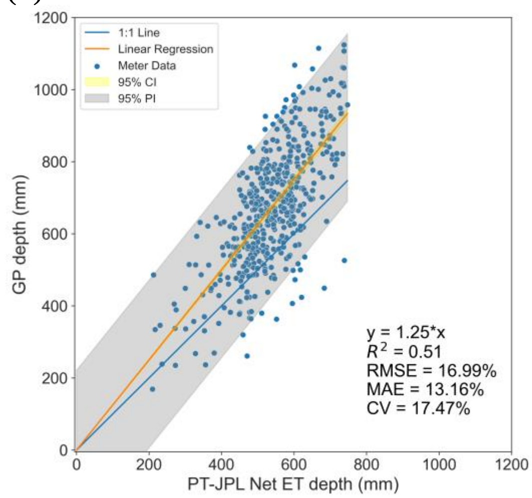
(a)



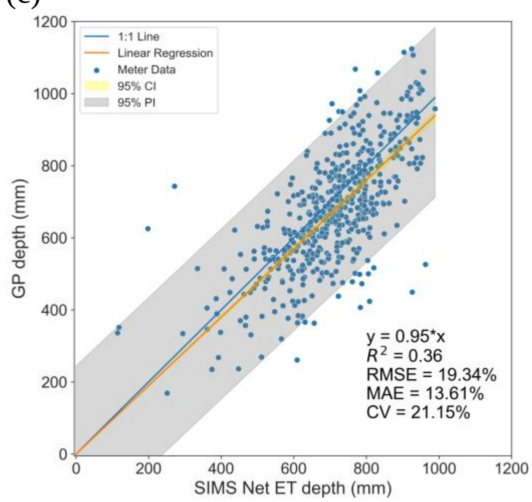
(b)



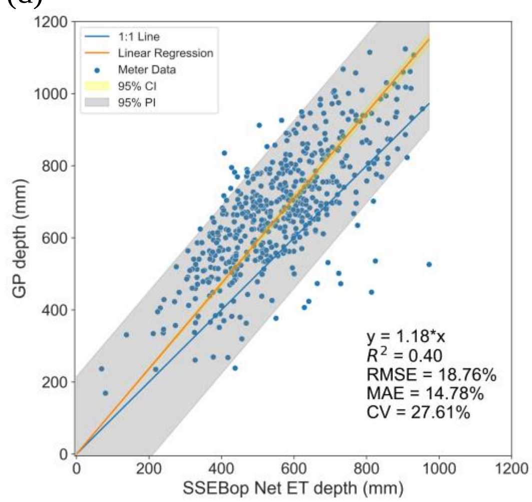
(c)



(d)

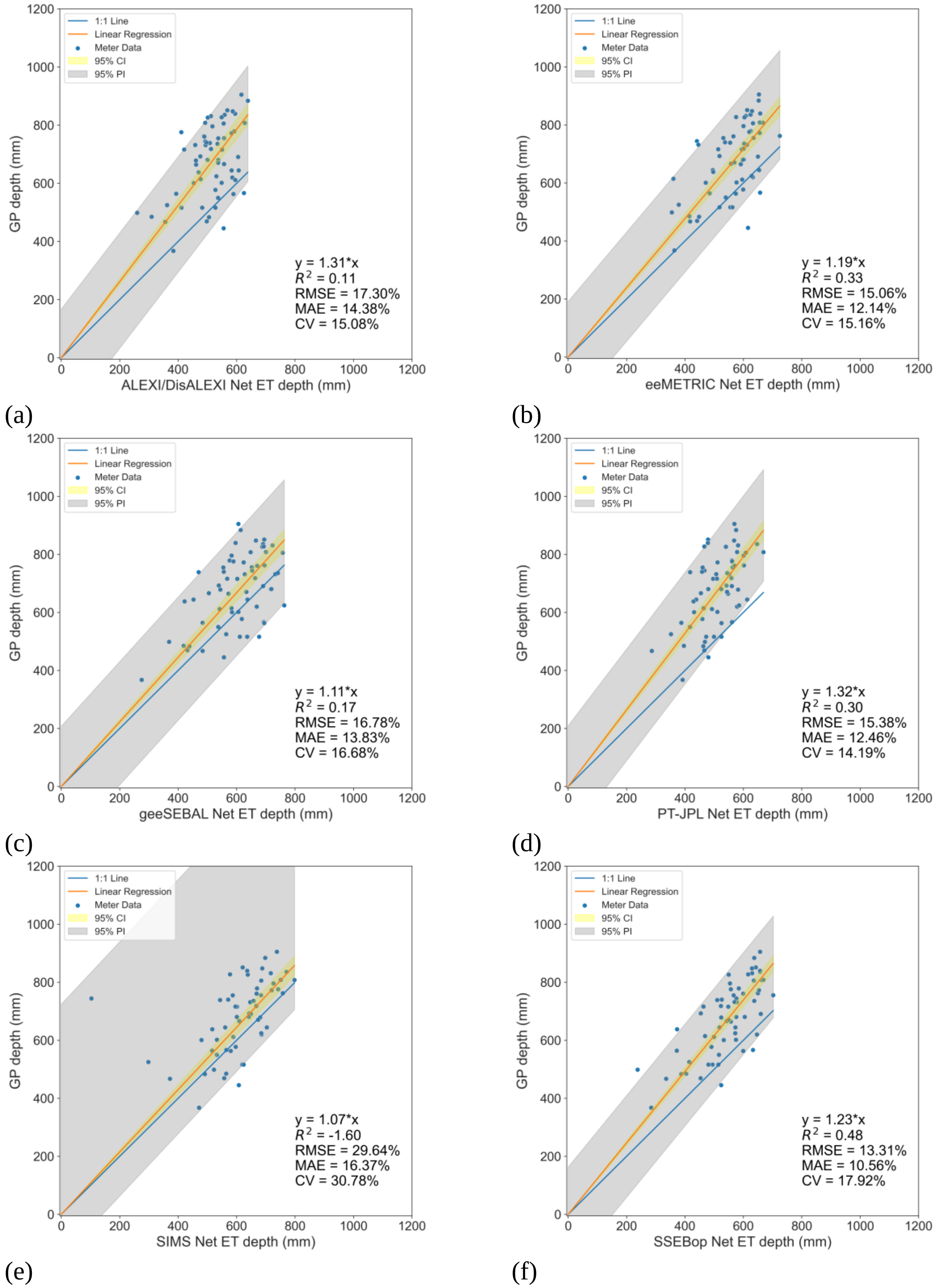


(e)

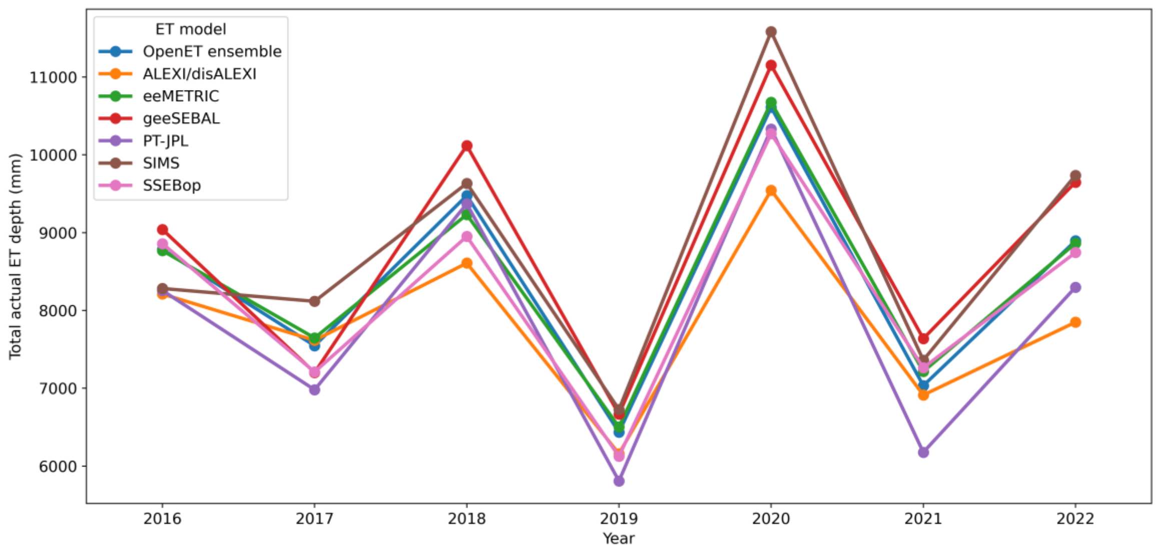


(f)

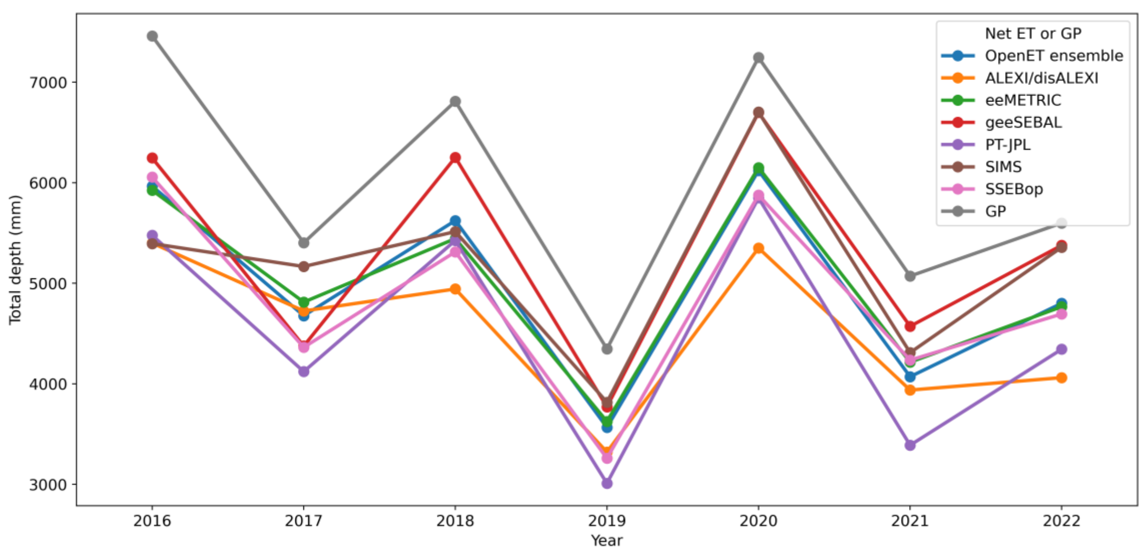
Supplementary Figure 3. Scatter plots of the linear regression models for (a) ALEXI/DisALEXI, (b) eeMETRIC, (c) geeSEBAL, (d) PT-JPL, (e) SIMS, and (f) SSEBop in DV, Nevada. The symbols and labels are defined in the main manuscript. The scatter plot of the OpenET ensemble is shown in Figure 7 (a) of the main manuscript.



Supplementary Figure 4. Scatter plots of the linear regression models for (a) ALEXI/DisALEXI, (b) eeMETRIC, (c) geeSEBAL, (d) PT-JPL, (e) SIMS, and (f) SSEBop in HB, Oregon. The symbols and labels are defined in the main manuscript. The scatter plot of the OpenET ensemble is shown in Figure 10 (a) of the main manuscript.



(a)



(b)

Supplementary Figure 5. Comparisons of the (a) total annual ET depths and (b) total $Net\ ET$ and total reported metered GP depths for each ET model in HB, Oregon.