

Knowledge-Guided Machine Learning Models to Upscale Evapotranspiration in the U.S. Midwest

Aleksei Rozanov

rozan012@umn.edu

Department of Geography, Environment and Society,

GEMS Informatics Center,

Department of Computer Science and Engineering

University of Minnesota - Twin Cities

Samikshya Subedi

subed036@umn.edu

Department of Soil, Water, and Climate,

GEMS Informatics Center

University of Minnesota - Twin Cities

Vasudha Sharma

vasudha@umn.edu

Department of Soil, Water, and Climate

University of Minnesota - Twin Cities

Bryan C. Runck*

runck014@umn.edu

GEMS Informatics Center

University of Minnesota - Twin Cities

*Corresponding Author

Abstract

Evapotranspiration (ET) plays a critical role in the land-atmosphere interactions, yet its accurate quantification across various spatiotemporal scales remains a challenge. *In situ measurement approaches*, like eddy covariance (EC) or weather station-based ET estimation, allow for measuring ET at a single location. Agricultural uses of ET require estimates for each field over broad areas, making it infeasible to deploy sensing systems at each location. This study integrates tree-based and knowledge-guided machine learning (ML) techniques with multispectral remote sensing data, gridded meteorology and EC data to upscale ET across the Midwest United States. We compare four tree-based models - Random Forest, CatBoost, XGBoost, LightGBM - and a simple feed-forward artificial neural network in combination with features engineered using knowledge-guided ML principles. Models were trained and tested on EC towers located in the Midwest of the United States using k-fold cross validation with k=5 and site-year, biome stratified train-test split to avoid data leakage. Results show that LightGBM with knowledge-guided features outperformed other methods with an $R^2=0.86$, $MSE=14.99 \text{ W}\cdot\text{m}^{-2}$ and $MAE = 8.82 \text{ W}\cdot\text{m}^{-2}$ according to grouped k-fold validation (k=5). Feature importance analysis shows that knowledge-guided features were most important for predicting evapotranspiration. Using the best performing model, we provide a data product at 500 m spatial and one-day temporal resolution for gridded ET for the period of 2019–2024. Intercomparison between the new gridded product and state-level weather station-based ET estimates show best-in-class correspondence.

Key Words

Evapotranspiration, agriculture, knowledge-guided machine learning, upscaling, MODIS, ERA5

Highlights

1. We developed a tree-based machine learning model combining remote sensing, meteorology, and eddy covariance observations to upscale evapotranspiration (ET) at 500 m daily resolution, achieving high accuracy ($R^2=0.86$, $MSE=14.99 \text{ W}\cdot\text{m}^{-2}$ and $MAE = 8.82 \text{ W}\cdot\text{m}^{-2}$). Our model integrates knowledge-guided machine learning by incorporating Penman-Monteith, demonstrating that physically based features significantly enhance ET prediction and dominate model feature importance rankings.
2. We generated a gridded ET dataset for the U.S. Midwest covering 2019–2024, offering consistent daily estimates suitable for agricultural applications, land surface modeling, and regional water management analysis.

Introduction

Evapotranspiration (ET) represents the largest component of the crop water balance and serves as a key variable in irrigation planning and water resource management (Hargreaves, 1985; Pereira, 1999; Yang, 2023). Actual evapotranspiration(ETa) is amount of total crop-water loss from specific crop and field whereas reference evapotranspiration (ETr or ETo) is the hypothetical water loss of tall grass alfalfa (ETr) or short grass (ETo) from a well-watered reference surface with no water stress (Allen et al., 1994). Crop coefficient (Kc) plays a crucial role in irrigation water management, as they reflect varying water needs of crops at different crop-growth stages(Pereira, 2021). Efficient irrigation management that leverages accurate ETa estimates can significantly optimize water use, conserve resources, minimize environmental impacts, and promote sustainable agriculture (Ray, 2024; Subedi et al., 2025). This approach aligns with the broader goal of computationally designed agricultural research that connects data-driven models directly to farmer and stakeholder needs (Kantar et al., 2025). Aligning water supply with crop demand supports not only productivity but also water availability, quality, and long-term economic viability.

Despite the critical importance of ET, accurately estimating it remains a challenge due to limited ground-based observations. Their spatial coverage is inherently limited. In contrast, remote sensing has emerged as a powerful tool to scale ET estimation over large areas. Satellite platforms, particularly those in polar orbits, provide consistent spatial and temporal coverage that enables regional-to-global monitoring. Nevertheless, capturing the dynamic and continuous behavior of ET as a process involving water vapor movement across heterogeneous landscapes remains a significant technical and methodological challenge (Senay, 2022).

Different methods have been developed to estimate evapotranspiration (ET), ranging from empirical and physically based equations to data-driven models. : 1) Direct measurement -Traditional methods, such as lysimeters and eddy covariance (EC) towers, have advanced our understanding of ETa at field scales (Allen et al., 1991; Baldocchi et al., 2001). 2) Remote sensing approaches to estimate ETa such as OpenET which includes different models like ALEXI/DIsALEXI, eeMETRIC, geeSEBAL, SIMS and SEEBoP (Anderson et al., 2007; Allen et al., 2007; Bastiaanssen et al., 1998; Pereira et al., 2020; Senay et al., 2013). 3) Weather based approaches like ETr/ ETo (Penman) and Hargreaves and Samani (HS) model, which requires weather variables to calculate reference evapotranspiration (Hargreaves & Samani, 1985; Allen et al., 1994).The Penman-Monteith (PM) equation is widely regarded as one of the most accurate and theoretically robust approaches. Recommended by the FAO-56 guidelines, the PM equation combines radiative, aerodynamic, and thermodynamic parameters to estimate reference evapotranspiration (ETo). Despite its accuracy, the PM method requires multiple meteorological inputs, which may not always be available in data-scarce regions. To address this, simpler empirical models like the HS model require only a limited number of inputs (primarily temperature and extraterrestrial radiation) making it more suitable for areas with sparse observational networks. ETo is calculated from the nearest available weather station and then scaled to field-level ETa using Kc values for the target crop and growth stage. The HS equation (Eq. 1) is expressed as (Moeletsi, 2013):

$$ET_{HS} = 0.135 \cdot K_T \cdot (T_{avg} + 17.8) \cdot \sqrt{T_{max} - T_{min}} \cdot R_a \cdot 0.408 \cdot d, \quad (1)$$

where ET_{HS} is estimated ET (mm/dekad), K_T – empirical coefficient , $T_{avg}, T_{max}, T_{min}$ are average, maximum, and minimum temperature ($^{\circ}C$), R_a – extraterrestrial radiation ($MJ \cdot m^{-2}$), dd – number of days in the dekad and 0.408 is a conversion factor from energy to water depth.

While convenient, these empirical methods must be calibrated and validated against local climatic data and conditions to ensure reliability (Hargreaves & Samani, 1985; Moeletsi, 2013). Errors in weather-based methods arise both from the representativeness of the weather station (e.g., distance from the field) and from generalized K_c values, which may not capture local management or environmental variability. Eddy covariance (EC) method is standard methods for measuring how energy and water move between land and atmosphere. What sets EC apart is that it measures sensible (H) and latent (LE) heat fluxes directly rather than estimating from other weather variables. Despite their values, EC towers are scattered sparsely. Upscaling ET_a measurements from EC towers to regional scales provides a more complete picture of energy and water exchanges, which helps us to understand how these systems respond to water and energy exchanges across the US Midwest and beyond (Jung et al., 2019; Xu et al., 2018).

Machine Learning Approaches to Estimate ET_a

In recent years, machine learning (ML) has emerged as a powerful alternative for estimating ET_a, especially in data-limited environments. Unlike empirical models, ML algorithms do not rely on predefined functional forms. Instead, they learn complex and often nonlinear relationships from historical data, making them well-suited for modeling highly variable hydrometeorological processes.

ML models such as Multilayer Perceptrons (MLPs), Extreme Gradient Boosting (XGBoost), Light Gradient Boosting Machine (LightGBM), and CatBoost have demonstrated competitive or superior performance compared to traditional ET models across various climates (Liang et al., 2023; Prokhorenkova et al., 2018; Ke et al., 2017; Chen & Guestrin, 2016). A key advantage of ML methods is their ability to make accurate predictions even when some input variables are missing or incomplete, which is particularly valuable in operational settings with inconsistent meteorological records. For example, Malik et al. (2017) applied a Multilayer Perceptron to model monthly pan-evaporation in the Indian Central Himalayas. The MLPNN demonstrated strong generalization across two study sites and outperformed traditional climate-based models, particularly at higher elevations. This superior performance was attributed to the model's ability to capture nonlinear interactions among climatic variables, which are capabilities that are often limited in empirical approaches.

Another study by dos Santos, Robson Argolo, et al. (2024) evaluated the performance of 13 different ML algorithms for ET prediction across three distinct biomes, using remote sensing data from Landsat 8–9 and Sentinel-2 satellites. The study highlighted the sensitivity of model performance to both biome type and input data sources. It demonstrated that gradient-boosting models, especially CatBoost and LightGBM, performed consistently well across varying environmental conditions. The study also reinforced the value of multispectral imagery in characterizing vegetation and surface energy dynamics relevant to ET estimation.

While machine learning models have demonstrated remarkable performance in modeling complex environmental processes (including ET_a), they often operate as black boxes, lacking physical interpretability and generalizability across regions or our out-of-sample observations. To address these

limitations, recent research has explored Knowledge-Guided Machine Learning (KGML), which integrates domain knowledge from physical models with data-driven algorithms. There are three primary ways that domain knowledge is integrated into ML according to Karpatne et al., 2024: knowledge-guided learning, knowledge-guided architecture, and knowledge-guided pretraining. In knowledge-guided learning, domain knowledge is incorporated into the training algorithm (for example, by adding domain-based constraints or physics-inspired terms to the loss function) to steer the model towards scientifically consistent solutions. Knowledge-guided architecture means incorporating prior knowledge directly into the model's design – for instance, encoding known physical laws or invariances within the neural network structure itself. Finally, knowledge-guided pretraining leverages domain knowledge to initialize model parameters (often by pretraining on simulation data from scientific models before fine-tuning on real observations), thereby infusing scientific knowledge into the model. Summing up, KGML approaches combine the flexibility of machine learning with the scientific rigor of physically based equations to constrain or inform the learning process. This hybrid strategy enhances model interpretability, improves robustness in data-scarce regions, and leads to improved predictive accuracy (Karpatne et al., 2024; Runck et al., 2022).

This study aims to develop a high-resolution evapotranspiration (ET) modeling framework that integrates machine learning with knowledge-guided strategies using remote sensing, meteorological data, and eddy covariance (EC) observations. Our objectives are threefold: (1) to evaluate the performance of the KGML model using a rigorous grouped k-fold cross-validation approach across diverse ecosystem types in the U.S. Midwest; (2) to quantify the contribution of physical and remote sensing features to ET predictions through feature importance analysis; and (3) to generate a daily gridded ET product at 500 m resolution for the period of 2019–2024, offering a valuable dataset for agricultural management, hydrological modeling, and regional climate analysis.

Materials and Methods

Study Area

Our study focuses on the Central Midwest and Great Plains regions of the United States (Fig. 1), encompassing the states of Minnesota, Iowa, Wisconsin, North Dakota, South Dakota, Nebraska, Illinois, Indiana, Missouri, Michigan, and Kansas. This region forms the heart of U.S. agricultural production and has experienced a marked increase in irrigation use over the past two decades, as producers seek to expand cultivation into more marginal lands and mitigate risks associated with increasing drought frequency and climate variability. The climate across this area transitions from humid continental in the eastern and northern parts (e.g., Minnesota, Wisconsin, Michigan) to semi-arid in the Great Plains (e.g., Nebraska, Kansas, the Dakotas). These gradients contribute to substantial variability in evapotranspiration driven by differences in precipitation, temperature, wind speed, and solar radiation. The region experiences warm to hot summers, with peak ET rates occurring during the growing season from May through August. The study period spans from 2000 to 2024, covering a wide range of hydroclimatic conditions, including major drought events and increasing trends in summer heat extremes, making this dataset suitable for developing and validating robust, scalable ET_a estimation models.

FLUXNET and AmeriFlux sites in the Central Midwest and Great Plains Region USA

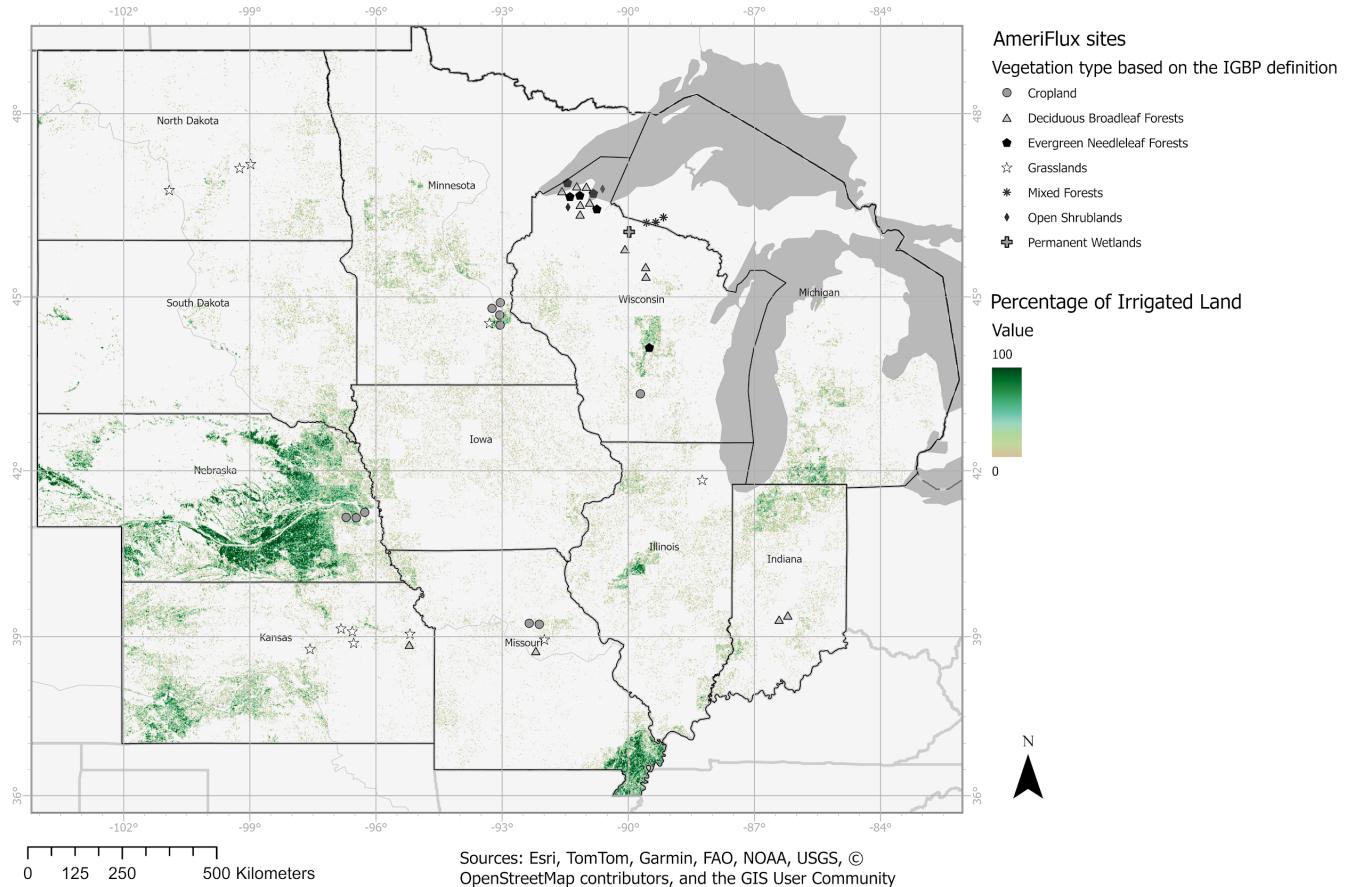


Figure 1. An overview of the Central Midwest and Great Plains Region USA, where the selected FLUXNET and AmeriFlux sites are located. To add more context to the study region, the percentage of irrigated land derived from MIRAD-US dataset by Pervez, Md Shahriar, and Jesslyn F. Brown. (2010) is also shown.

Data

To solve the ET upscaling problem we developed a unique dataset consisting of four data sources: ERA5-Land (Muñoz Sabater, 2019), MOD09GA v061 (Vermote and Wolfe, 2021), AmeriFlux and FLUXNET (Pastorello et al., 2020). AmeriFlux and FLUXNET provide ground-based measurements of evapotranspiration and related fluxes using eddy covariance towers across a range of ecosystems.

ERA5-Land retrospective climate reanalysis is a global gridded dataset produced using 4dVAR data assimilation (Courtier, 1994) techniques providing information about terrestrial meteorology from 1950 onwards. In the current work, we used daily aggregates of ERA5-Land distributed by Google Earth Engine (Gorelic et al., 2017). The selected meteorology variables reflect the primary physical processes governing ET:

1. Air temperature and dewpoint temperature at 2 m control vapor pressure deficit, a key driver of atmospheric demand for ET.
2. Wind components and surface pressure influence turbulent transfer and enhance vapor removal at the surface, and affects air density, which modulates latent heat fluxes, respectively.
3. Surface net solar radiation directly supplies the energy needed for evaporation and transpiration.
4. Total evaporation and total precipitation represent a process-based model proxy for moisture fluxes, and water availability, setting boundary conditions for soil evaporation and plant transpiration respectively.

Since daily ET is not only defined by the current weather conditions, but also the meteorological events of the recent past (Kabala, 2025), all the meteorologies were derived as 30 days time series, where the last day in the array is the actual day of ET observation and other 29 characterise the weather in the recent past.

MOD09GA v061 provides estimates of surface spectral reflectance from the Moderate Resolution Imaging Spectroradiometer (MODIS) as it would appear at ground level without atmospheric scattering or absorption effects. The dataset is distributed as a daily gridded Level 2G product in a sinusoidal projection, offering 500 m reflectance data along with 1 km observation and geolocation statistics. Similar to ERA5-Land, MOD09GA v061 was accessed through Google Earth Engine (GEE). To extract proxies potentially related to ET, we included the following variables:

1. Sensor and viewing geometry (sensor zenith angle, sensor azimuth angle, solar zenith angle, solar azimuth angle), providing information about observation geometry and solar energy input.
2. Surface reflectance bands 1–7 (Red, NIR1, Blue, Green, NIR2, SWIR1, SWIR2), capturing vegetation and land surface conditions closely tied to ET processes.
3. Clouds QA (State 1km), which was binary decoded to flag pixels affected by clouds or shadows (1) versus clear observations (0).

Lastly, the target (dependent) variable for the supervised ML task was derived from AmeriFlux and FLUXNET eddy covariance (EC) time series at 38 sites (Fig. 1; Appendix A and B) across the U.S. Midwest, processed using the standardized OneFlux methodology (Pastorello et al., 2020). Specifically, we used LE_F_MDS (latent energy flux) as the dependent variable, pre-filtered to include only daily values with $\text{LE_F_MDS_QC} \geq 0.75$, ensuring that observations with less than 25% gap-filling were retained.

Since evapotranspiration is directly related to latent energy (LE) through a simple energy-to-depth conversion ($\text{ET} = \text{LE} / \lambda$, where λ is the latent heat of vaporization) (Pereira, 1998), LE serves as a physically consistent and robust proxy for ET in this study. Additionally, site latitude, longitude, IGBP (International Geosphere-Biosphere Programme; Loveland and Belward, 1997) ecosystem classification and observation day of year were extracted from the AmeriFlux and FLUXNET products and used as features in our own dataset.

Feature Engineering

To enhance the predictive power of the future models, feature engineering is a step as important as feature selection. Different machine learning models vary in their ability to internally capture complex mathematical relationships. These relationships include, but are not limited to, counts, standard deviations, distances, and powers. If a model can learn such patterns inherently, manually crafting those features may be unnecessary, but according to (Heaton, 2016), many classical ML architectures struggle to capture certain relationships.

In the current work, we used three different approaches for feature engineering and extraction: First, we constructed weather-based time series features. Since ERA5-Land reanalysis was derived as a 30 days time series, this data contains temporal information that cannot be directly inferred by classical tree-based models. To extract the temporal information about meteorology, we computed minimum, maximum, and standard deviation values for each ERA5 variable in the dataset. For variables representing cumulative meteorological characteristics (e.g. surface net solar radiation, total evaporation, and total precipitation), we computed 7 and 30 days rolling sums, and for other variables, rolling means. These steps resulted in a time series dataset where each meteorological time series was transformed into a vector with the following components: last value, minimum, maximum, standard deviation, the rolling 30-day average or sum, and the rolling 7-day average or sum. Then all the vectors were stacked into one 1-dimensional vector.

Second, we constructed remote sensing-based features. Previous work (Running, 2004; Xiao, 2004; Rozanov and Gribanov, 2023) has shown that remote sensing vegetation indexes can be a strong predictor of terrestrial fluxes. In our current work, we used Normalized Difference Vegetation Index (NDVI; Eq. 2), Enhanced Vegetation Index (EVI; Eq. 3), Green Normalized Difference Vegetation Index (GNDVI; Eq. 4), Soil Adjusted Vegetation Index (SAVI; Eq. 5), Atmospherically Resistant Vegetation Index (ARVI; Eq. 6).

$$NDVI = \frac{NIR - Red}{NIR + Red}, \quad (2)$$

$$EVI = \frac{NIR - Red}{NIR + 6 Red - 7.5 Blue + 1}, \quad (3)$$

$$GNDVI = \frac{NIR - Green}{NIR + Green}, \quad (4)$$

$$SAVI = \frac{NIR - Red}{NIR + Red + 0.5} 1. 5, \quad (5)$$

$$ARVI = \frac{NIR - 2Red + Blue}{NIR + 2 Red - Blue}, \quad (6)$$

Lastly, constructed KGML features. In the context of ET upscaling, although the PM equation (Eq. 7) could be incorporated as a custom loss term, such an approach risks diverging from the observed measurements, given that PM does not fully capture the true variability of ET across different ecosystems

(Choudhury, 1997). To mitigate this limitation of PM, while still leveraging prior physical knowledge, we incorporated PM-derived ET estimates computed from ERA5-Land meteorological features as input features to the machine learning models. Specifically, six new features were generated: the last recorded value, minimum, maximum, standard deviation, 30-day rolling mean, and 7-day rolling mean of PM-based ET time series.

$$ET_o = \frac{0.408\Delta(R_n - G) + \gamma \frac{900}{T+273} U_2(e_s - e_a)}{\Delta + \gamma(1 + 0.34U_2)}, \quad (7)$$

where R_n is net radiation at the crop surface ($\text{MJ} \cdot \text{m}^{-2} \cdot \text{day}^{-1}$), G is soil heat flux density ($\text{MJ} \cdot \text{m}^{-2} \cdot \text{day}^{-1}$), T is air temperature at 2 m height ($^{\circ}\text{C}$), U_2 is wind speed at 2 m height ($\text{m} \cdot \text{s}^{-1}$), e_s is saturation vapor pressure (kPa), e_a is Actual vapor pressure (kPa), Δ is slope of the vapor pressure curve ($\text{kPa} \cdot ^{\circ}\text{C}^{-1}$), and γ is psychrometric constant ($\text{kPa} \cdot ^{\circ}\text{C}^{-1}$).

Model Description

In the current work we attempted to use several recently developed tree-based ML architectures, such as Extreme Gradient Boosting (XGBoost; Chen et al., 2016), Categorical Boosting (CatBoost; Prokhorenkova et al., 2018) and LightGBM (Ke et al., 2017) against commonly-used baselines, Random Forest (RF; Breiman, 2001) and Artificial Neural Network (ANN), also known as Multilayer Perceptron (MLP; Rosenblatt, 1958). Fig. 2 gives a high-level overview of our methodology, including data acquisition, pre-processing and modeling itself. In the following, we describe each model and our implementation. Across all tree-based models, we selected optimal hyperparameters through grid search, unless otherwise noted.

First, RF is an ensemble of regression trees trained using a bootstrapping technique and was implemented with the scikit-learn (Pedregosa, 2011; ver. 1.6.1) Python package. Grid search resulted in 100 trees (`n_estimators=100`) and a maximum tree depth of 22 (`max_depth=22`).

Second, CatBoost is a gradient boosting algorithm designed to handle categorical features natively, employing ordered boosting to prevent target leakage and improve prediction stability. The model was implemented using the CatBoost (ver. 1.2.7) Python library with tree depth equal to 16, number of iterations set to 2000, and bootstrapping type equal to Poisson.

Third, XGBoost is a gradient boosting framework that builds additive tree models sequentially. It was trained with the following hyperparameters: a learning rate of 0.1, a maximum tree depth of 14, a minimum child weight of 5, and a gamma value of 0.6 to regularize tree splitting. Additionally, column subsampling (`colsample_bytree=0.9`) and row subsampling (`subsample=0.8`) were applied to enhance model generalization. The ensemble consisted of 100 trees (`n_estimators=100`). The model was trained using the `xgboost (3.0.2)` python package.

Four, LightGBM is a gradient boosting framework (ver. 4.6.0) based on decision tree algorithms, optimized for efficiency with histogram-based learning and leaf-wise tree growth. The model was tuned with 80 leaves, 0.05 learning rate, 0.7 colsample_bytree, optimizing RMSE.

Lastly, we built an MLP model with two hidden layers of 400 and 100 neurons and a tanh activation function applied to a weighted linear combination of each neuron using pytorch (Paszke, 2019; 2.6.0). Batch normalization and dropout with probability of 0.3 were used as well as regularization techniques to prevent overfitting.

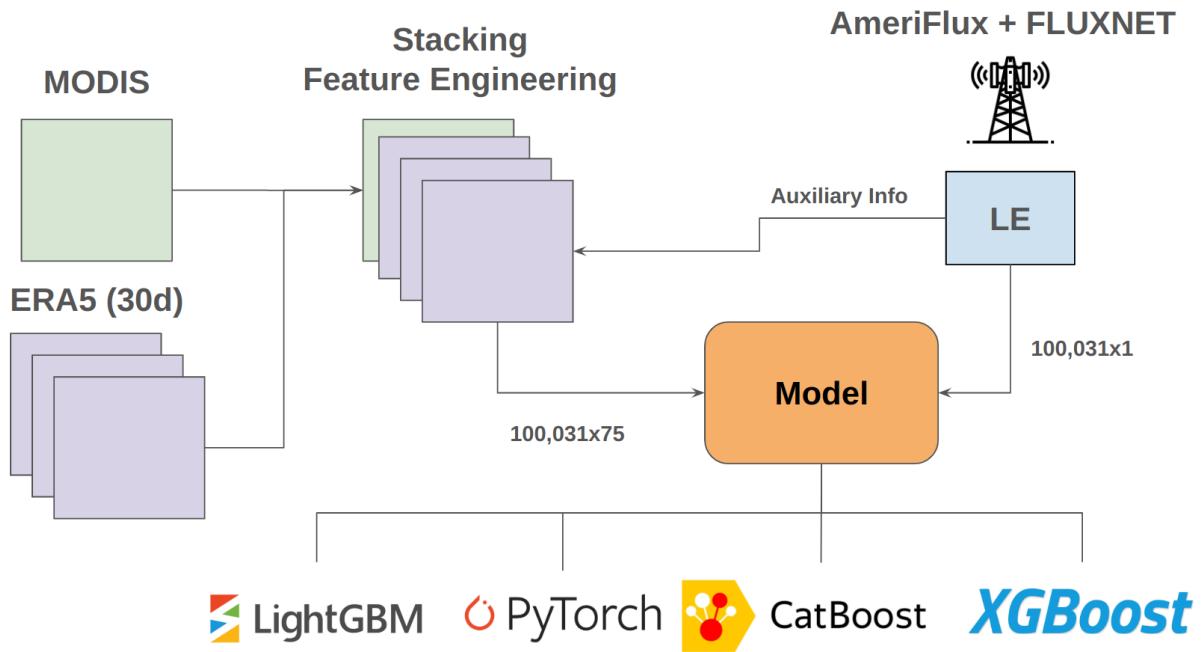


Figure 2. A data flow diagram, demonstrating the developed framework with data sources and processing stages applied to them.

Results and Discussion

Evaluation Metrics Selection

To compare models performance three common metrics were used: mean absolute error (**MAE**; Eq. 8), Root Mean Square Error (**RMSE**; Eq. 9), and **R²** (Eq. 10):

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_{i,pred} - y_{i,true}| , \quad (8)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_{i,pred} - y_{i,true})^2}, \quad (9)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_{i,pred} - \bar{y}_{true})^2}{\sum_{i=1}^n (y_{i,true} - \bar{y}_{true})^2}, \quad (10)$$

where n is the number of data samples, y_{pred} is a model's prediction and y_{true} is the resembling ground-true value.

Model Performance

To ensure rigorous model evaluation and prevent data leakage across spatial and temporal domains, we used GroupKFold ($k=10$) cross-validation from *scikit-learn* python library with site-year as a grouping variable. This approach ensures that all observations from a given site and year are confined to either the training or validation set, never both. Such partitioning prevents information from leaking across temporal or spatial boundaries, which is particularly important in flux upscaling tasks where environmental variables exhibit strong site-level and seasonal autocorrelations. Traditional train-test splits, even when stratified by biome (IGBP), risk overestimating performance by allowing similar conditions from the same site or time period to influence both training and testing. In contrast, our group-based cross-validation yields higher error than train-test split but provides a more realistic estimate of generalization performance, simulating the model's ability to extrapolate to entirely unseen site-year conditions. We discarded all other validation strategies that did not explicitly account for grouping, as they can produce biased metrics and inflate confidence in model predictions.

Model performance is summarized in Table 1. LightGBM outperformed all other models achieving the lowest RMSE = 14.99 W·m⁻² (0.861 mm day⁻¹) and MAE = 8.82 W·m⁻² (0.528 mm day⁻¹), and the highest R² = 0.86 during the cross validation. CatBoost and XGBoost demonstrated similar performance, yet worse than LightGBM, whereas RF and ANN had the weakest performance on the given task.

Table 1. Comparison of the trained models on the test dataset from the site-year split.

Model	Metric		
	R²	RMSE, W·m⁻²	MAE, W·m⁻²
Random Forest	0.838	16.07	9.585
CatBoost	0.832	16.374	9.629

XGBoost	0.840	15.985	9.46
LightGBM	0.860	15.990	8.823
ANN	0.756	19.720	13.444

In addition to global metrics, we evaluated model performance across individual months and IGBP land cover classes (Fig. 3) to assess its temporal and ecological generalization capabilities. Monthly error analysis showed that May, June, and July, corresponding to the period of highest ET flux, exhibited the largest errors, with RMSE values of $22.07 \pm 0.49 \text{ W} \cdot \text{m}^{-2}$ ($0.775 \pm 0.017 \text{ mm day}^{-1}$), $25.85 \pm 1.09 \text{ W} \cdot \text{m}^{-2}$ ($0.913 \pm 0.039 \text{ mm day}^{-1}$), and $23.58 \pm 0.36 \text{ W} \cdot \text{m}^{-2}$ ($0.835 \pm 0.012 \text{ mm day}^{-1}$) respectively (\pm indicates the standard error across folds). In contrast, winter months (December, January, and February) showed substantially lower errors and variability, with RMSEs of $3.86 \pm 0.265 \text{ W} \cdot \text{m}^{-2}$ ($0.133 \pm 0.009 \text{ mm day}^{-1}$), $3.8 \pm 0.053 \text{ W} \cdot \text{m}^{-2}$ ($0.132 \pm 0.002 \text{ mm day}^{-1}$), and $5.06 \pm 0.14 \text{ W} \cdot \text{m}^{-2}$ ($0.175 \pm 0.005 \text{ mm day}^{-1}$), respectively.

When stratified by IGBP land cover class, the model predicted ET in Mixed Forests (MF) most accurately, with an RMSE of $6.462 \pm 0.69 \text{ W} \cdot \text{m}^{-2}$ ($0.227 \pm 0.024 \text{ mm day}^{-1}$). Higher errors were observed in Croplands (CRO), Wetlands (WET), and Grasslands (GRA), which had RMSE values of $16.09 \pm 0.75 \text{ W} \cdot \text{m}^{-2}$ ($0.556 \pm 0.027 \text{ mm day}^{-1}$), $17.718 \pm 0.33 \text{ W} \cdot \text{m}^{-2}$ ($0.622 \pm 0.012 \text{ mm day}^{-1}$), and $16.34 \pm 0.374 \text{ W} \cdot \text{m}^{-2}$ ($0.576 \pm 0.013 \text{ mm day}^{-1}$), respectively. These elevated errors likely reflect greater heterogeneity in canopy structure and soil moisture dynamics within these ecosystems.

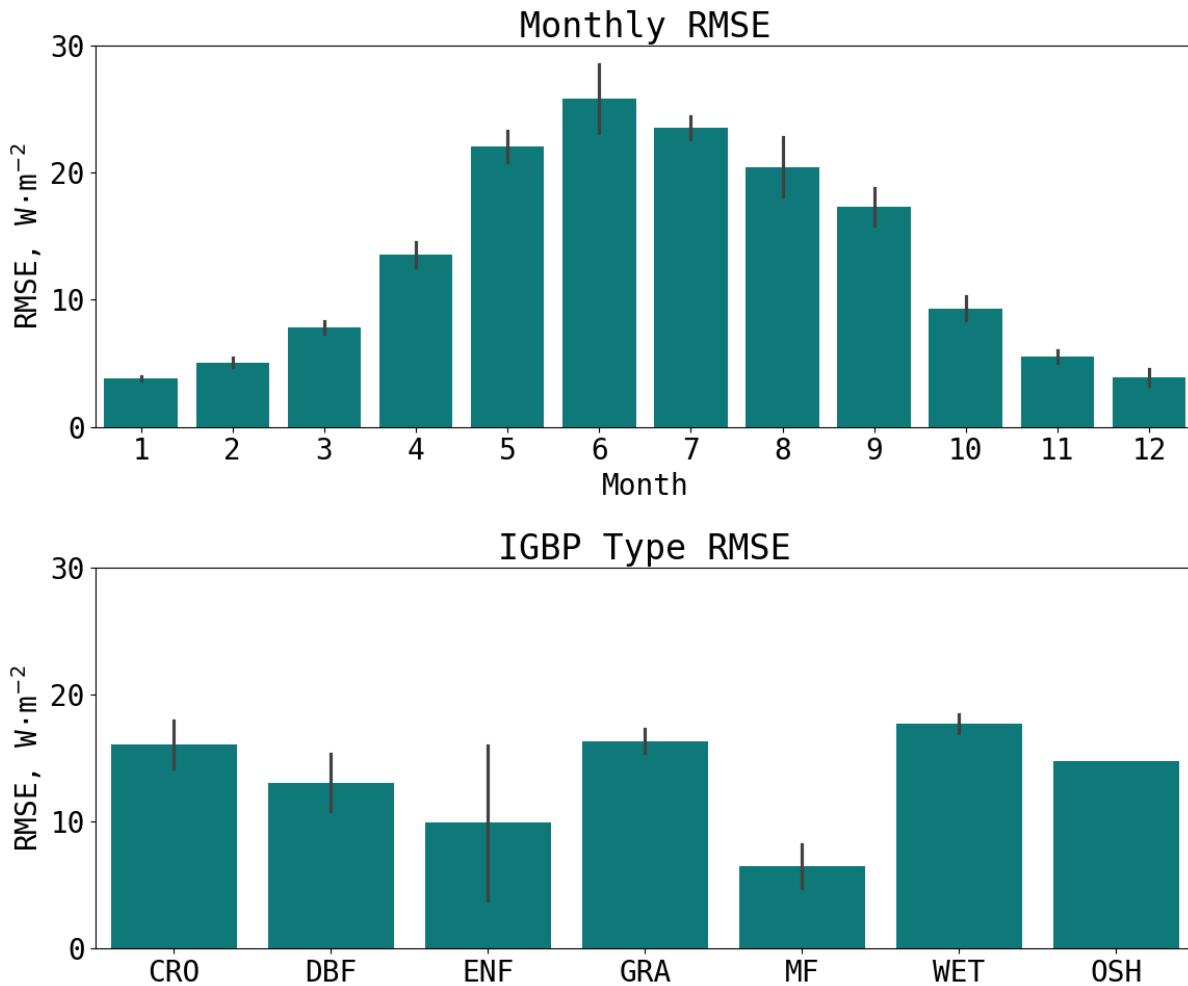


Figure 3. Error by month and IGBP type. CRO, DBF, ENF, GRA, MF, WET and OSH represent Croplands, Deciduous Broadleaf Forests, Evergreen Needleleaf Forests, Grasslands, Mixed Forests, Wetlands and Open Shrublands, respectively.

Lastly, we conducted a feature importance analysis based on gain using the native LightGBM tools (Fig. 4). Feature importance by gain quantifies the contribution of each feature to reducing the overall loss, specifically RMSE, by aggregating the improvements made at each split where the feature is used across all trees. The results align with established understanding of the primary drivers of LE (and ET), highlighting the critical role of historical radiation, air temperature, and water balance in shaping ET dynamics across sites. Notably, the most important feature was the knowledge-guided input derived from the Penman-Monteith equation, demonstrating that the model strongly leverages physically based prior knowledge. This finding directly supports one of the core objectives of this study: to evaluate whether incorporating physical process knowledge enhances model performance. The results demonstrate that the most accurate model consistently prioritized the physically informed feature when making ET predictions.

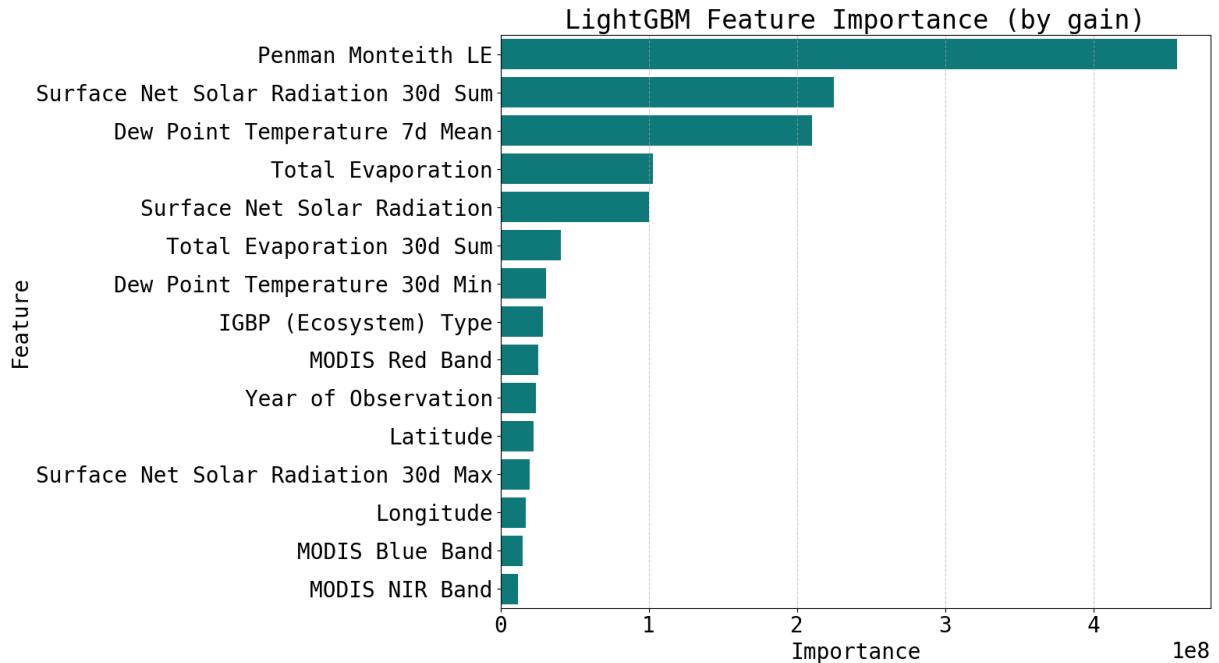


Figure 4. Feature importance analysis performed for the LightGBM model, where importances are computed based on the information gain at each split which is based on the given feature. The top 15 most impactful features are shown.

Data Product

To support the needs of decision-makers in the U.S. Upper Midwest, we release an open-access dataset (Fig. 5) generated by our framework for the period 2019–2024. The data have a spatiotemporal resolution of 500 m and 1 day, covering latitudes 36°N to 49°N and longitudes –104°W to –82°W. The dataset is distributed in NetCDF4 format.

To ensure the quality of the generated data, we performed an additional evaluation by comparing monthly aggregated values with (i) the OpenET ensemble median product (Melton et al., 2021) obtained via Google Earth Engine (OpenET/ENSEMBLE/CONUS/GRIDMET/MONTHLY/v2_0), and (ii) Penman–Monteith ET_r estimates derived from ground-based meteorological observations at 13 Minnesota Mesonet sites. The results (Fig. 6) show strong agreement with OpenET ($r = 0.94$, RMSE = 15.24 mm month^{–1}), reasonable agreement with the PM-based approach ($r = 0.89$, RMSE = 35.52 mm month^{–1}), and tighter correspondence with the PM-based approach than OpenET ($r = 0.88$; RMSE = 34.9 mm month^{–1}) a finding that is useful for local decision-makers evaluating data products for decision-support.

Average Upscaled Evapotranspiration, 2019-2024

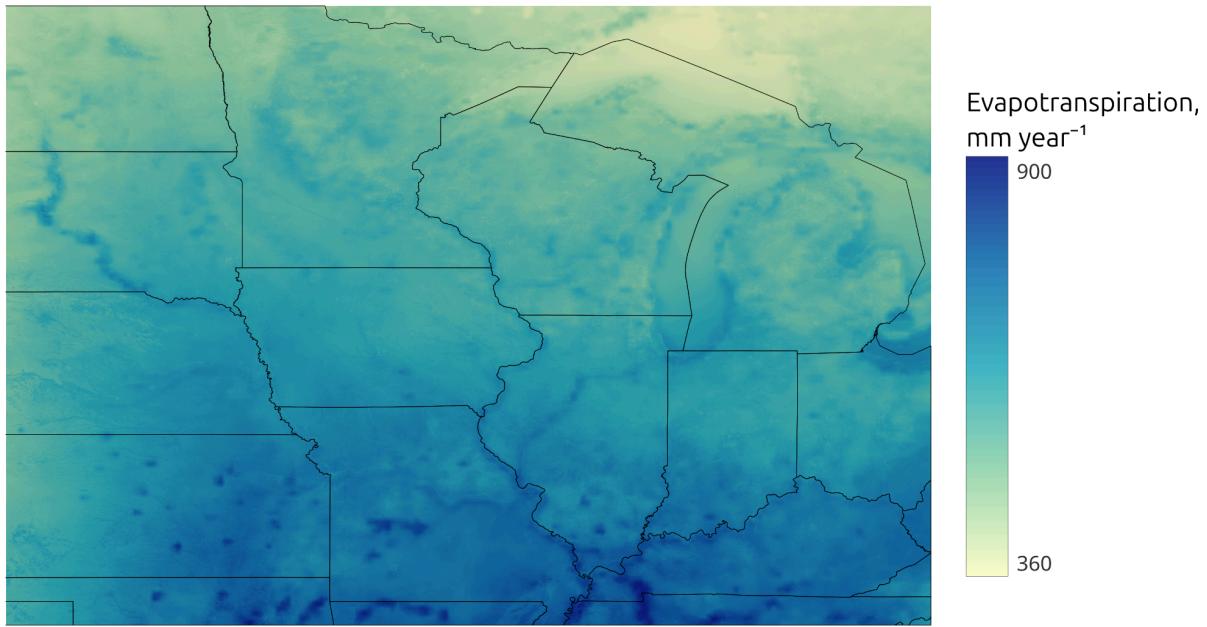


Figure 5. Average annual ET between 2019 and 2024 derived from the upscaled data product.

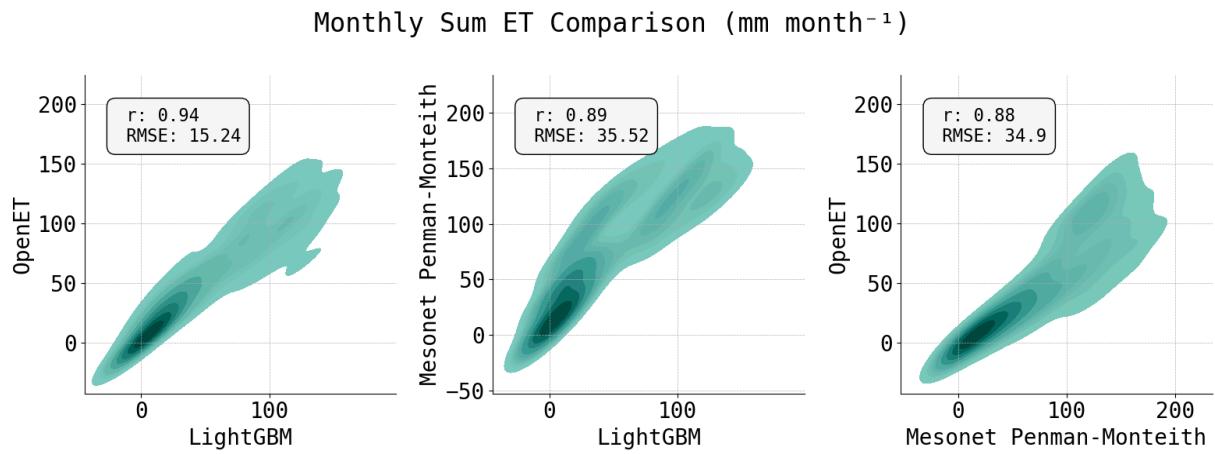


Figure 6. Intercomparison of monthly sum ET (mm month^{-1}) with ensemble median from OpenET and PM-derived ET based on the Mesonet stations meteorologies for 13 sites in Minnesota, US.

Conclusion

Here, we presented a high-resolution (500 m, daily) ET product for the U.S. Upper Midwest, generated using a LightGBM-based machine learning model informed by physical principles. The model was rigorously validated using GroupKFold cross-validation by site-year, ensuring realistic assessments of generalization performance and minimizing spatial and temporal leakage—an often-overlooked issue in environmental upscaling. Among evaluated methods, LightGBM demonstrated superior predictive accuracy across sites, seasons, and land cover types, with the highest R^2 (0.855) and lowest RMSE (15.2 $\text{W}\cdot\text{m}^{-2}$). Model performance declined during peak ET months and over heterogeneous land cover types, consistent with increased flux variability.

Our results underscore the value of hybrid approaches: the model's top feature was a knowledge-guided input derived from the Penman–Monteith equation, highlighting the synergistic benefit of combining process understanding with data-driven methods. This demonstrates that physically informed features can meaningfully improve performance and interpretability in machine learning-based flux modeling.

To support practical applications, we publicly release the full ET dataset for 2019–2024 as an open-access resource: <https://doi.org/10.13020/37XE-QQ18>. The product shows strong agreement with independent benchmarks, including OpenET and ground-based PM estimates, confirming its utility for operational and research use. The dataset can inform irrigation planning, drought monitoring, and hydrological modeling across the Midwest.

Additionally, to promote transparency, open-source science, and research reproducibility, we make available all Python code developed during this study in a public Git repository: https://github.com/RTGS-Lab/ET_LCCMR. This includes data preprocessing workflows, model training scripts, and evaluation routines, enabling other researchers to build on our work and adapt the framework to new regions, fluxes, or remote sensing inputs.

CRediT Authorship Contribution Statement

Aleksei Rozanov: Conceptualization, Methodology, Formal analysis, Data curation, Writing – original draft, review & editing, Visualization. **Samikshya Subedi:** Writing – original draft, review & editing. **Vasudha Sharma:** Writing – review& editing. **Bryan Runck:** Writing – review & editing, Conceptualization, Resources, Methodology, Project administration.

AI Usage Statement

Portions of this manuscript, including language refinement and clarity improvements, were assisted by OpenAI's ChatGPT. All content was reviewed and edited by the authors to ensure accuracy and originality.

Declaration of competing interests

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

Acknowledgements

Funding for this project was provided by the Minnesota Environment and Natural Resources Trust Fund as recommended by the Legislative-Citizen Commission on Minnesota Resources (LCCMR) ENRTF 2021-266.

Data Availability

The dataset is openly available at the Data Repository for the University of Minnesota (DRUM): <https://conservancy.umn.edu/items/ada05f34-e8a6-492f-83dd-04db7983ac19>. The code created during this study is available through GitHub: https://github.com/RTGS-Lab/ET_LCCMR.

References

- Allen, R.G., Howell, T.A., Pruitt, W.O., Walter, I.A., & Jensen, M.E. (1991). Lysimeters for evapotranspiration and environmental measurements. ASCE.
- Anderson, M. C., Norman, J. M., Mecikalski, J. R., Otkin, J. A., & Kustas, W. P. (2007). A climatological study of evapotranspiration and moisture stress across the continental United States based on thermal remote sensing: 1. Model formulation. Journal of Geophysical Research: Atmospheres, 112(D10). <https://doi.org/10.1029/2006JD007506>
- Baldocchi, D., Falge, E., Gu, L., Olson, R., Hollinger, D., Running, S., et al. (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), 2415-2434. [https://doi.org/10.1175/1520-0477\(2001\)082<2415:FANTTS>2.3.CO;2](https://doi.org/10.1175/1520-0477(2001)082<2415:FANTTS>2.3.CO;2)
- Bastiaanssen, W. G., Menenti, M., Feddes, R. A., & Holtslag, A. A. M. (1998). A remote sensing surface energy balance algorithm for land (SEBAL). 1. Formulation. Journal of hydrology, 212, 198-212. [https://doi.org/10.1016/S0022-1694\(98\)00253-4](https://doi.org/10.1016/S0022-1694(98)00253-4).
- Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32. <https://doi.org/10.1023/A:1010933404324>.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). <https://doi.org/10.1145/2939672.2939785>

Choudhury, B.J. (1997). Global pattern of potential evaporation calculated from the Penman-Monteith equation using satellite and assimilated data. *Remote Sensing of Environment*, 61(1), 64-81. [https://doi.org/10.1016/S0034-4257\(96\)00241-6](https://doi.org/10.1016/S0034-4257(96)00241-6)

Courtier, P., Thépaut, J.N., & Hollingsworth, A. (1994). A strategy for operational implementation of 4D-Var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, 120(519), 1367-1387. <https://doi.org/10.1002/qj.49712051912>

Kantar, M., Ewing, P., Bančić, J., Fumia, N., Garba, I., Jamshidi, S., ... & Wang, D. (2025). Computational design for more engaged, impactful, and dynamic agricultural research. *Crop Science*, 65(2), e70034.

dos Santos, R.A., Zeri, M., de Faria, C.M.A., et al. (2024). Improving actual evapotranspiration estimates through an integrated remote sensing and cutting-edge machine learning approach. *Computers and Electronics in Agriculture*, 225, 109258. <https://doi.org/10.1016/j.compag.2024.109258>

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., & Moore, R. (2017). Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202, 18-27. <https://doi.org/10.1016/j.rse.2017.06.031>

Hargreaves, G.H., & Samani, Z.A. (1985). Reference crop evapotranspiration from temperature. *Applied Engineering in Agriculture*, 1(2), 96-99. <https://doi.org/10.13031/2013.26773>.

Heaton, J. (2016). An empirical analysis of feature engineering for predictive modeling. In SoutheastCon 2016 (pp. 1-6). IEEE. <https://doi.org/10.1109/SECON.2016.7506650>.

Jung, M., Koirala, S., Weber, U., Ichii, K., Gans, F., Camps-Valls, G., ... & Reichstein, M. (2019). The FLUXCOM ensemble of global land-atmosphere energy fluxes. *Scientific data*, 6(1), 74. <https://doi.org/10.1038/s41597-019-0076-8>

Kabala, J. P., Sobrino, J. A., Crisafulli, V., Skoković, D., & Battipaglia, G. (2025). geeSSEBI: Evaluating Actual Evapotranspiration Estimated with a Google Earth Engine Implementation of S-SEBI. *Remote Sensing*, 17(3), 395. <https://doi.org/10.3390/rs17030395>.

Karpatne, A., Jia, X., & Kumar, V. (2024). Knowledge-guided machine learning: Current trends and future prospects. arXiv preprint arXiv:2403.15989. <https://doi.org/10.48550/arXiv.2403.15989>

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T.Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.

Liang, Y., Feng, D., Sun, Z., & Zhu, Y. (2023). Evaluation of empirical equations and machine learning models for daily reference evapotranspiration prediction using public weather forecasts. *Water*, 15(22), 3954. <https://doi.org/10.3390/w15223954>

Liu, L., Xu, S., Tang, J., Guan, K., Griffis, T.J., Erickson, M.D., et al. (2022). KGML-ag: a modeling framework of knowledge-guided machine learning to simulate agroecosystems: a case study of estimating N₂O emission using data from mesocosm experiments. *Geoscientific Model Development*, 15(7), 2839-2858. <https://doi.org/10.5194/gmd-15-2839-2022>

Loveland, T.R., & Belward, A.S. (1997). The international geosphere biosphere programme data and information system global land cover data set (DISCover). *Acta Astronautica*, 41(4-10), 681-689. [https://doi.org/10.1016/S0094-5765\(98\)00050-2](https://doi.org/10.1016/S0094-5765(98)00050-2)

Malik, A., Kumar, A., & Kisi, O. (2017). Monthly pan-evaporation estimation in Indian central Himalayas using different heuristic approaches and climate based models. *Computers and Electronics in Agriculture*, 143, 302-313. <https://doi.org/10.1016/j.compag.2017.11.008>

Melton, F., Huntington, J., Grimm, R., Herring, J., Hall, M., Rollison, D., Erickson, T., Allen, R., Anderson, M., Fisher, J., Kilic, A., Senay, G., Volk, J., Hain, C., Johnson, L., Ruhoff, A., Blankenau, P., Bromley, M., Carrara, W., Daudert, B., Doherty, C., Dunkerly, C., Friedrichs, M., Guzman, A., Halverson, G., Hansen, J., Harding, J., Kang, Y., Ketchum, D., Minor, B., Morton, C., Revelle, P., Ortega-Salazar, S., Ott, T., Ozdogon, M., Schull, M., Wang, T., Yang, Y., & Anderson, R. (2021). OpenET: Filling a critical data gap in water management for the western United States. *Journal of the American Water Resources Association*. <https://doi.org/10.1111/1752-1688.12956>

Miralles, D.G., et al. (2016). The WACMOS-ET project—Part 2: Evaluation of global terrestrial evaporation data sets. *Hydrology and Earth System Sciences*, 20(2), 823-842. <https://doi.org/10.5194/hess-20-823-2016>.

Moeletsi, M.E., Walker, S., & Hamandawana, H. (2013). Comparison of the Hargreaves and Samani equation and the Thornthwaite equation for estimating dekadal evapotranspiration in the Free State Province, South Africa. *Physics and Chemistry of the Earth*, 66, 4-15. <https://doi.org/10.1016/j.pce.2013.08.003>

Muñoz Sabater, J. (2019). ERA5-Land hourly data from 1950 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). <https://doi.org/10.24381/cds.e2161bac>

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. *arXiv preprint*, arXiv:1912.01703. <https://doi.org/10.48550/arXiv.1912.01703>

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.W., ... & Law, B. (2020). The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Scientific Data*, 7(1), 225. <https://doi.org/10.1038/s41597-020-0534-3>

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.

Pereira, L.S., Perrier, A., Allen, R.G., & Alves, I. (1999). Evapotranspiration: concepts and future trends. *Journal of Irrigation and Drainage Engineering*, 125(2), 45-51.
[https://doi.org/10.1061/\(ASCE\)0733-9437\(1999\)125:2\(45\)](https://doi.org/10.1061/(ASCE)0733-9437(1999)125:2(45))

Pereira, L. S., Paredes, P., Melton, F., Johnson, L., Wang, T., López-Urrea, R., ... & Allen, R. G. (2020). Prediction of crop coefficients from fraction of ground cover and height. Background and validation using ground and remote sensing data. *Agricultural Water Management*, 241, 106197.
<https://doi.org/10.1016/j.agwat.2020.106197>

Pervez, M.S., & Brown, J.F. (2010). Mapping irrigated lands at 250-m scale by merging MODIS data and national agricultural statistics. *Remote Sensing*, 2(10), 2388-2412. <https://doi.org/10.3390/rs2102388>.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, 31.

Ray, S., & Majumder, S. (2024). Water management in agriculture: Innovations for efficient irrigation. *Modern Agronomy*, 169-185.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386. <https://doi.org/10.1037/h0042519>.

Rozanov, A.P., & Gribanov, K.G. (2023). A neural network model for estimating carbon fluxes in forest ecosystems from remote sensing data. *Atmospheric and Oceanic Optics*, 36(4), 323-328.
<https://doi.org/10.1134/S1024856023040152>

Runck, B. C., Joglekar, A., Silverstein, K. A., Chan-Kang, C., Pardey, P. G., & Wilgenbusch, J. C. (2022). Digital agriculture platforms: Driving data-enabled agricultural innovation in a world fraught with privacy and security concerns. *Agronomy journal*, 114(5), 2635-2643.

Running, S.W., Nemani, R.R., Heinsch, F.A., Zhao, M., Reeves, M., & Hashimoto, H. (2004). A continuous satellite-derived measure of global terrestrial primary production. *BioScience*, 54(6), 547-560.
[https://doi.org/10.1641/0006-3568\(2004\)054\[0547:ACSMOG\]2.0.CO;2](https://doi.org/10.1641/0006-3568(2004)054[0547:ACSMOG]2.0.CO;2)

Senay, G. B., Bohms, S., Singh, R. K., Gowda, P. H., Velpuri, N. M., Alemu, H., & Verdin, J. P. (2013). Operational evapotranspiration mapping using remote sensing and weather datasets: A new parameterization for the SSEB approach. *JAWRA Journal of the American Water Resources Association*, 49(3), 577-591. <https://doi.org/10.1111/jawr.12057>

Senay, G.B., Friedrichs, M., Morton, C., Parrish, G.E., Schauer, M., Khand, K., ... & Huntington, J. (2022). Mapping actual evapotranspiration using Landsat for the conterminous United States: Google Earth Engine implementation and assessment of the SSEBop model. *Remote Sensing of Environment*, 275, 113011. <https://doi.org/10.1016/j.rse.2022.113011>

Subedi, S., Kechchour, A., Kantar, M., Sharma, V., & Runck, B. C. (2025). Can gridded real-time weather data match direct ground observations for irrigation decision-support?. *Agrosystems, Geosciences & Environment*, 8(2), e70100.

Vermote, E., & Wolfe, R. (2021). MODIS/Terra Surface Reflectance Daily L2G Global 1km and 500m SIN Grid V061 [Data set]. NASA Land Processes Distributed Active Archive Center. <https://doi.org/10.5067/MODIS/MOD09GA.061>

Xiao, X., Hollinger, D., Aber, J., Goltz, M., Davidson, E.A., Zhang, Q., & Moore III, B. (2004). Satellite-based modeling of gross primary production in an evergreen needleleaf forest. *Remote Sensing of Environment*, 89(4), 519-534. <https://doi.org/10.1016/j.rse.2003.11.008>

Xu, T., Guo, Z., Liu, S., He, X., Meng, Y., Xu, Z., ... & Song, L. (2018). Evaluating different machine learning methods for upscaling evapotranspiration from flux towers to the regional scale. *Journal of Geophysical Research: Atmospheres*, 123(16), 8674-8690. <https://doi.org/10.1029/2018JD028447>

Yang, Y., Roderick, M.L., Guo, H., Miralles, D.G., Zhang, L., Fatichi, S., ... & Yang, D. (2023). Evapotranspiration on a greening Earth. *Nature Reviews Earth & Environment*, 4(9), 626-641. <https://doi.org/10.1038/s43017-023-00464-3>

Supplemental Materials

Appendix A. FLUXNET sites used in this study.

Site	Link	Latitude	Longitude
US-IB2	https://fluxnet.org/sites/siteinfo/US-IB2	41.8406	-88.241
US-Los	https://fluxnet.org/sites/siteinfo/US-Los	46.0827	-89.9792
US-MMS	https://fluxnet.org/sites/siteinfo/US-MMS	39.3232	-86.4131
US-Ne1	https://fluxnet.org/sites/siteinfo/US-Ne1	41.1651	-96.4766
US-Ne2	https://fluxnet.org/sites/siteinfo/US-Ne2	41.1649	-96.4701
US-Ne3	https://fluxnet.org/sites/siteinfo/US-Ne3	41.1797	-96.4397
US-Syv	https://fluxnet.org/sites/siteinfo/US-Syv	46.242	-89.3477
US-WCr	https://fluxnet.org/sites/siteinfo/US-WCr	45.8059	-90.0799
US-Wi0	https://fluxnet.org/sites/siteinfo/US-Wi0	46.6188	-91.0814
US-Wi1	https://fluxnet.org/sites/siteinfo/US-Wi1	46.7305	-91.2329
US-Wi2	https://fluxnet.org/sites/siteinfo/US-Wi2	46.6869	-91.1528
US-Wi3	https://fluxnet.org/sites/siteinfo/US-Wi3	46.6347	-91.0987
US-Wi4	https://fluxnet.org/sites/siteinfo/US-Wi4	46.7393	-91.1663
US-Wi6	https://fluxnet.org/sites/siteinfo/US-Wi6	46.6249	-91.2982
US-Wi8	https://fluxnet.org/sites/siteinfo/US-Wi8	46.7223	-91.2524

Appendix B. AmeriFlux sites used in this study.

Site	Link	Latitude	Longitude
US-CS2	https://ameriflux.lbl.gov/sites/siteinfo/US-CS2	44.1467	-89.5002

US-DFC	https://ameriflux.lbl.gov/sites/siteinfo/US-DFC	43.3448	-89.7117
US-KFS	https://ameriflux.lbl.gov/sites/siteinfo/US-KFS	39.0561	-95.1907
US-KLS	https://ameriflux.lbl.gov/sites/siteinfo/US-KLS	38.7745	-97.5684
US-Kon	https://ameriflux.lbl.gov/sites/siteinfo/US-Kon	39.0824	-96.5603
US-MMS	https://ameriflux.lbl.gov/sites/siteinfo/US-MMS	39.3232	-86.4131
US-MO1	https://ameriflux.lbl.gov/sites/siteinfo/US-MO1	39.2298	-92.1167
US-MO2	https://ameriflux.lbl.gov/sites/siteinfo/US-MO2	38.9488	-91.9945
US-MO3	https://ameriflux.lbl.gov/sites/siteinfo/US-MO3	39.2322	-92.1493
US-MOz	https://ameriflux.lbl.gov/sites/siteinfo/US-MOz	38.7441	-92.2000
US-Ro1	https://ameriflux.lbl.gov/sites/siteinfo/US-Ro1	44.7143	-93.0898
US-Ro2	https://ameriflux.lbl.gov/sites/siteinfo/US-Ro2	44.7288	-93.0888
US-Ro4	https://ameriflux.lbl.gov/sites/siteinfo/US-Ro4	44.6781	-93.0723
US-Ro5	https://ameriflux.lbl.gov/sites/siteinfo/US-Ro5	44.691	-93.0576
US-Ro6	https://ameriflux.lbl.gov/sites/siteinfo/US-Ro6	44.6946	-93.0578
US-Syy	https://ameriflux.lbl.gov/sites/siteinfo/US-Syy	46.2420	-89.3477
US-Wi0	https://ameriflux.lbl.gov/sites/siteinfo/US-Wi0	46.6188	-91.0814
US-Wi1	https://ameriflux.lbl.gov/sites/siteinfo/US-Wi1	46.7305	-91.2329
US-Wi3	https://ameriflux.lbl.gov/sites/siteinfo/US-Wi3	46.6347	-91.0987
US-Wi4	https://ameriflux.lbl.gov/sites/siteinfo/US-Wi4	46.7393	-91.1663
US-Wi6	https://ameriflux.lbl.gov/sites/siteinfo/US-Wi6	46.6249	-91.2982
US-Wi8	https://ameriflux.lbl.gov/sites/siteinfo/US-Wi8	46.7223	-91.2524

US-xDC	https://ameriflux.lbl.gov/sites/siteinfo/US-xDC	47.1617	-99.1066
US-xKA	https://ameriflux.lbl.gov/sites/siteinfo/US-xKA	39.1104	-96.6129
US-xKZ	https://ameriflux.lbl.gov/sites/siteinfo/US-xKZ	39.1008	-96.5631
US-xNG	https://ameriflux.lbl.gov/sites/siteinfo/US-xNG	46.7697	-100.9154
US-xST	https://ameriflux.lbl.gov/sites/siteinfo/US-xST	45.5089	-89.5864
US-xTR	https://ameriflux.lbl.gov/sites/siteinfo/US-xTR	45.4937	-89.5857
US-xUK	https://ameriflux.lbl.gov/sites/siteinfo/US-xUK	39.0404	-95.1921
US-xUN	https://ameriflux.lbl.gov/sites/siteinfo/US-xUN	46.2339	-89.5373
US-xWD	https://ameriflux.lbl.gov/sites/siteinfo/US-xWD	47.1282	-99.2414
