Upscaling Global Hourly GPP with Temporal Fusion Transformer(TFT)

John Calzaretta*
Mary Chau*
Rumi Nakagawa*
john.calzaretta@berkeley.edu
marychau@berkeley.edu
ruminakagawa@berkeley.edu
University of California, Berkeley
Berkeley, CA, USA

ABSTRACT

With increasing attention and investments in carbon reduction initiatives from both public and private sectors, reliable estimates of carbon flux, including Gross Primary Productivity (GPP), are crucial for assessing the accountability and effectiveness of climate change initiatives. FLUXNET provides empirical estimates of GPP based on the measured Net Ecosystem Exchange (NEE) in flux towers. However, global estimates of GPP continue to have high uncertainty, due to the limited number of flux sites that provide CO_2 fluxes. Some machine learning models applied to GPP upscaling had limited performance due to lack to predictor features at higher temporal resolutions and irregularity in the number and distributions of missing values. With the focus on developing a global upscaling model for hourly GPP, this research explored a novel application of Temporal Fusion Transformer (TFT) on time-series problem without the availability of past target values. Model development was supplemented by Random Forest Regressor (RFR) and XGBoost (XGB), followed by the hybrid model of TFT and tree algorithms. As a result, one of the developed models outperformed the latest study by 2.9% improvement in NSE (70.4%) and 10.2% of improvement in RMSE (3.54). Additionally, this study utilize interpretable outputs of TFT to analyze model results and the temporal dynamics of feature importance per prediction. The study lays the foundation for future utilization of TFT in addressing generic upscaling problems across various fields.

KEYWORDS

GPP, Upscaling, Temporal Fusion Transformer(TFT), Random Forest Regressor, XGBoost

*All three authors contributed equally to this research.

Authors' address: John Calzaretta, john.calzaretta@berkeley.edu; Mary Chau, marychau@berkeley.edu; Rumi Nakagawa, ruminakagawa@berkeley.edu, University of California, Berkeley, 2200 University Ave Ste 1500, Berkeley, CA, USA, 94720.

0004-5411/2023/0-ART0 $0.00
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
John Calzaretta, Mary Chau, and Rumi Nakagawa. 2023. Upscaling Global Hourly GPP with Temporal Fusion Transformer(TFT). *J. ACM* 0, 0, Article 0 (2023), 24 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Despite the temporary slowdown during the 2020 pandemic, the persistent rise in CO_2 emissions has prompted an unprecedented mobilization of government agencies, industries, and individuals to address climate change. In 2021, the United States government pledged a USD 369 billion investment for sustainable energy and climate change initiatives. The European Union, a pioneer of green policies, committed to reducing greenhouse gas emissions to 55% of 1990 levels by 2030. The private sector has also been actively investing in renewable energy and the carbon market. For example, Alphabet, Google's parent company, supports the equivalent of 5.5 GW of renewable energy worldwide. As investments in carbon reduction initiatives continue to grow, reliable estimation of carbon flux is essential for informed green policies and global carbon budgets, as well as for ensuring the effectiveness and accountability of these actions.

Carbon dioxide exchanges are measured by flux towers, which are tall, instrumented structures that measure the exchange of gasses, such as carbon dioxide, between the Earth's surface and atmosphere. FLUXNET is a global network of regional networks that ties together earth system scientists to share flux data and methods. There are over 1,000 active and historic flux measurement sites globally, but only around 260 flux towers has been used for measuring carbon flux, resulting in available data being geographically sparse. Due to the limited number of locations with actual GPP measurements, upscaling is necessary to estimate global GPP. Upscaling refers to the process of extrapolating GPP measurements from available flux data to global regions where no flux sites exist. This upscaling task is also challenging due to regional biases of flux data as the majority of operational sites are located in North America and Europe. Improving empirical models for global inference with state-of-the-art machine learning algorithms will enhance the universal accessibility of quantified CO_2, regardless of the availability of flux tower equipment.

FLUXNET provides the following primary carbon-related features: 1) Net Ecosystem Exchange (NEE): The net CO_2 exchange that flux tower measures. 2) Ecosystem Respiration (RECO): An estimated carbon respiration. 3) Gross Primary Productivity (GPP):

A measure of carbon sequestration by vegetation. It is obtained by the difference of NEE and estimated RECO. This study focuses on providing a global estimate of GPP, as NEE is comprised of two opposing dynamics that offset each other, thereby contributing to the improvement of NEE quality.

Past literature has studied the upscaling of global GPP products at various temporal resolutions, such as annual [3], monthly[12][14], daily[27] and half-hourly[5]level. One of the primary challenges when modeling at a higher temporal resolution was the lack of features availability at the same temporal resolution. In previous studies, meteorological features were typically limited to daily resolution as the highest level of granularity. Additionally, due to the computational resources needed for time-aware models (such as LSTM), many machine learning studies in this field have not accounted for the temporal dynamics within the data. When the available resource is limited, trade-off between the processing time, and limitation in the size of train period or the number of applicable features are inevitable.

On the flip side, predicting GPP in lower temporal resolution also has constraints due to the metabolism and physiological responses of plants. First, some vegetation types have a large variance of GPP in a single day. Additionally, the diurnal cycles of GPP also vary across seasons. Another challenge is that the actual flux tower measurement times vary by site. Since flux tower GPP typically shows low or even negative values during the nighttime, aggregating to lower resolutions can lead to a disconnection from the actual phenomenon. Therefore, the lower the resolution of GPP aggregation, the more diurnal information is lost from the site. These ambiguous aggregated GPP values may also be too noisy for the model to learn the real world phenomenon.

The Temporal Fusion Transformer (TFT) model has several key characteristics that appear promising in overcoming these limitations; 1) TFT works datasets of time-series data with various duration and capable of preforming forecast on new entities that are unseen before. Such capability is crucial because flux tower observation duration varies from decades to just several weeks, and to enable upscaling estimation, the model must be able preform predictions on areas that were not in its training data. 2) The model takes in heterogeneous time-series data inputs, including addition to input features from time-varying continuous and categorical variables to static metadata, and thus is allowed to take the full advantages of the data available. 3) TFT utilizes a combination of LSTM and self-attention mechanism to learn historical patterns in both short and long-term. 4) Lastly, This model is also highly interpretable, and enables insights into features importances and attention fluctuations across time, which can collectively enable deeper ecological insights. 5) TFT architecture is acclaimed for its efficiency of computational resources [17], and is thus better equipped to work with data sets with at higher resolution.

This study aimed to apply Temporal Fusion Transformer (TFT) to hourly global GPP upscaling by incorporating time-aware elements into the solution with two objectives; improving on model performances and analyze temporal dynamic of influential features via TFT's interpretable model outputs. A TFT model of the ideal upscaling scenario, where past GPP measurements were presented for all areas, were implemented to benchmark the best possible TFT performance. For real-life upscaling application where not past GPP measurement were available, Random Forest Regressor (RFR), XGBoost (XGB) models were used to establish baseline model performance, followed by two TFT modeling approaches: 1) Modeling with no past GPP measurement as feature input, and 2) Two-stage modeling with using estimated past GPP measurements from the predicted GPP values from either of the tree models. Result analysis focused on the TFT's explainability outputs and was able to derive a new analytic approach by breaking down features significance by encoder time steps that provide insights into the temporal influences of the features for model predictions.

2 LITERATURE REVIEW

Historically, the first global product of GPP was provided by the MOD17 MODIS project [23][38]. MOD17 is part of the NASA Earth Observation System (EOS) program, which is the first satellite-driven dataset to monitor vegetation productivity on a global scale. MOD17 algorithm applies original radiation efficiency logic that Monteith developed in 1972. The theoretical model suggests that productivity of annual crops under well-watered and fertilized conditions is linearly related to the amount of absorbed solar energy. The latest version (MOD17A2H Version 6.1) provides data since 2000, with 8-day, monthly, and annual temporal resolution. The FLUXNET project was founded in 1997, originally with the support from NASA to develop accessible ground truth. It was followed by the cooperation of the scientific community of flux scientists to develop networks across Euroflux (Europe), AmeriFlux (North America) and AsiaFlux (Asia).

The first empirical approach that applied machine learning to achieve upscaling GPP was the study that applied ANNs (Artificial neural networks) in European forests [21]. Support vector regression models[36][28][10] as well as ensemble tree models[33][34][12][5] have been actively studied algorithms. Tramontana compared sixteen machine learning algorithms, including kernel methods, neural networks, tree methods and regression splines[27]. The research group found that the performance of models were highly consistent among different machine learning algorithms or experimental setups, while the performance showed site-dependencies.

In regards to time series forecasting with deep learning algorithms, upscaling global GPP is an unexplored field. LSTM(Long Short Term Memory) was applied to build a forecasting model of NEE with remote sensing, climate and eddy-covariance flux datasets[4]. Since the study of NEE inference focused on the memory effects of sites, the forecasting was made on sites which were equivalent to the sites in the the training set.

Combinations of predictor variables vary by studies. When a study scopes upscaling GPP, satellite-based remote sensing data are commonly applied due to the wide availability in local sites [33][34][32]. Climate data[13]][14], temperature, and water availability[15] are also applied for modeling. Such predictor variables are categorized as meteorological variables[27][5], and it is another core components of predictor variables. As remote sensing features, Tramontana applied Land Surface Temperature (LST)[29][30], Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI)[9], Leaf Area Index (LAI), fraction of absorbed photosynthetic active radiation (fPAR)[19], BRDF-corrected surface reflectances [26], Normalized Difference Water Index (NDWI)[7]

and the Land Surface Water Index (LSWI)[35]. Air temperature (Tair), global radiation (Rg), vapor pressure deficit (VPD), and precipitation were applied as meteorological features.

Along with the development of available technology for modelings, the temporal resolution for the upscaling has become higher through the past two decades. Empirical upscaling of monthly average was continuously studied with a tree ensemble approach[12][14] and median annual GPP was studied between the two researches by Jung[3]. Daily resolution was studied in 2016[27], followed by the research of half-hourly resolution in 2018[5].

Bodesheim built a predictive upscaling model of GPP in half-hourly temporal resolution, using Random Forest Regressor (RFR) because of its fast learning and testing capabilities. Besides the PFT(Plant Functional Type), RS + METEO from the study by Tramontana in 2016[27] were applied as predictor variables. Mean seasonal cycles were used for some variables such as Normalized Difference Vegetation Indices (NDVI), Normalized Difference Water Index (NDWI), Land Surface Temperature (LST), and the fraction of Photosynthetically Active Radiation (fPAR). Most features that Bodesheim[5] applied had daily granularity as the highest temporal resolution, which meant that most features had unique values in each of 48 time points. Because of this limitation, the study applied two approaches. One was to build 48 independent models in each half-hour across 24 hours, and the other was to build a single model that contained values in any of the half hours of each day. The author added the first order temporal derivative to the model in the second approach, in order to express the uptrend or downtrend of the target variable at half-hourly level. Except the derivative variable, the potential radiation was the only variant feature in the half-hourly resolution. It indicated that the only two features were the substantial determinant of the half-hourly RFR model. Another limitation of the study was the availability of predictor variables in global sites. While meteorological variables, such as air temperature or vapor pressure deficit (VPD), contributed to improve the model performance, the availability of those variables were limited to flux tower sites. Second prediction approach with a single model had 0.67 efficiency in global sites. The performance increased to 0.71 with half-hourly meteorological data, however, the inference was only available in flux-tower sites. When the author experimented modeling with Mean Seasonal Cycle (MSC), it performed better than the half-hourly basis, potentially because of the varieties of determinants in the predictor variables.

To learn from and build upon past literature, our study attempts to improve GPP modeling performance by performing rigorous feature selection and also applying novel, time-aware GPP models. Additionally, this paper adds to the scientific understanding of carbon flux by contributing interpretable outputs from transformer models to enable deeper insights into the phenomena of GPP.

3 DATA SOURCES

The goal of upscaling is to predict GPP based on spatially sparse ground truth observations using global remote-sensing-derived and meteorological products. To accomplish this task effectively, a wide set of meaningful predictor features must be considered and evaluated. Taking advantage of available data types and the flexibility around input features of the TFT model, this study uses numerical, categorical, and static data as model inputs. The numerical features include data of various resolutions including monthly, 16-days, 8-days, 4-days, daily, and hourly levels. As global upscale requires features remain available even in locations with no flex tower, the only flux-tower-site-dependent feature was the GPP measurements (i.e. the target variable) from FLUXNET. Rest of the features were independent from the flux tower sites.

Note the features used in this study did not include the sites' geolocation (e.g. latitude and longitude) and year of the records to prevent model from overfitting on a target area's geolocation and the carbon fertilization effect [cite?] respectively. Omitting those features allows the study to focus on temporal effects of the meteorological and remote sensing data.

3.1 FLUXNET: Ground Truth GPP

FLUXNET is an international network that ties regional networks of earth system scientists. Eddy covariance, the standard method to measure gas fluxes between biosphere and atmosphere, is applied to measure the cycling of carbon, water, and energy. Eddy covariance is the standard method that ecosystem scientists use to measure gas fluxes between ecosystems and atmosphere. FLUXNET2015 is the latest FLUXNET data, which is hosted by Lawrence Berkeley National Laboratory. Eddy covariance method offers a mean of sequentially observed net fluxes between land and atmosphere [2][1] Not only the net ecosystem CO_2 exchange(NEE), but also water vapor and energy exchange are observed.

In FLUXNET, Gross Primary Production (GPP) is calculated by taking the difference between NEE and estimated Ecosystem Respiration (RECO). There are two types of GPP provided in FLUXNET: the nighttime based approach, and the daytime based approach. The nighttime based approach uses nighttime data to parameterize a respiration model[22] that is then applied to the whole dataset to estimate RECO. The measured time directly affects the estimation of RECO, which affects indirectly to GPP values. In this study, nighttime based approach is chosen. NEE is filtered with two different USTAR thresholds, which leads to two distinct hourly GPPs. According to FLUXNET, the USTAR thresholds are applied to remove the NEE values to avoid having false emission pulses due to accumulated CO_2 under the canopy, which is not detected by the storage system of the tower. FLUXNET provides Constant USTAR Threshold (CUT) and Variable USTAR Threshold (VUT). VUT is applied in this research, therefore USTAR thresholds vary by years.

While flux tower GPP is obtained from the empirically observed actual NEE, some GPP values are negative. This is because of the distinct methodologies between the observation of NEE and the estimation of the RECO. It is one of the limitations of the currently available technology to observe best possible GPP. Removing negative GPP leads to potential bias that may affect the learned patterns of the model, therefore any preprocessing for negative GPP is not conducted in this study.

The available hourly GPP records in FLUXNET have two main characteristics; 1) Most of the flux towers are concentrated in North America and Europe, making the data biased to those locations. 2) As shown in Figure 1, the duration of observation from flux towers not only varies drastically, from over 20 years to just a couple of months, but many also contain large gaps between observed

Figure 1: The timelines of available observations of flux tower sites by land-cover type (IGBP)

periods. These inconsistency and irregularities of the temporal data make time-aware modeling difficult, since most time-series models require continuous inputs. Having larger amount of gaps in time steps increases the dependency on gap-filled time steps which may result in the higher deviation from the original data.

3.2 Global Product: Remote-Sensing Data

3.2.1 MODIS. MODIS (Moderate Resolution Imaging Spectroradiometer) is the key instrument for Terra (EOS AM-1) and Aqua (EOS PM-1) satellites. According to NASA, both MODISs view the surface of the Earth each day or each of two days, and they acquire data in 36 spectral bands or groups of wavelengths. The data obtained by MODIS satisfies spatial resolution by 1 km or even better[16].

Following features are applied in this study; 1) MCD43C4 (Daily resolution): NIRv (Near-Infrared Reflectance of Vegetation[1]), NDVI (Normalized Difference Vegetation Index), EVI (Environmental Vulnerability Index), NDWI (Normalized Difference Water Index), percentage of snow, PET (Evapotranspiration), Surface reflectance b1 – b7 3) MCD15A3H (4-day, 8-day resolution): LAI (Surface reflectance Band 7 (SWIR3)), fPAR (Fraction of Photosynthetically Active Radiation) 4) MCD12Q1 (Annual resolution): PFT (Plant functional type) 5) MYD11A1 (Daily resolution): Daytime LST (Land Surface Temperature), Nighttime LST(Land Surface Temperature) 6) MODIS IGBP TOBE UPDATED

Tramontana used quality assurance/quality control (QA/QC) in MODIS to identify the quality of pixels. The author assigned the actual value when the 25% of the pixels had good quality at the time of the snapshot, and kept them blank if they did not reach the threshold. On the contrary, our study did not make use of QC for two reasons; 1. Replacing values with blank (zero) for time with low quality pixels may mislead the model that the actual value were zero. 2. The TFT model did not allow having missing values in any of the feature, therefore the imputation was required. Assuming that existing values was better than to fill with values from imputation logic, this study kept the original values instead of the replacement.

3.3 Global Product: Remote-Sensing-Derived Data

3.3.1 CSIF. Solar induced chlorophyll fluorescence (SIF) from space is an electromagnetic signal emitted by the chlorophyll of plants. It monitors the photosynthetic activity of terrestrial ecosystems to address the limitation of SIF in inconsistent footprint of measurements, low spatial and temporal resolution, and the high uncertainty of the individual retrievals. Contiguous SIF (CSIF) is trained with surface reflectance of MODIS and SIF from Orbiting Carbon Observatory-2 (OCO-2) through a neural network [37]. CSIF-SIFdaily (all-sky daily average SIF) is applied in this study with 4-day resolution.

3.3.2 BESS_Rad. Breathing Earth System Simulator (BESS) is a process-based simple model that couples atmosphere and canopy radiative transfers, canopy photosynthesis, transpiration, and energy balance. Following the original publication [24][11], BESS shortwave radiation, PAR and diffuse PAR products are studied [25]. Having MODIS atmospheric GPP product as an input, those values are predicted by radiative transfer model and ANN (artificial neural network). Therefore, using features from BESS in this research implies the indirect potential impact of theoretically estimated conventional GPP. The dataset is provided in daily temporal resolution, and BESS PAR, diffuse BESS PAR, BESS RSDN[2] are applied for this study.

3.4 Global Product: Meteorological Data

3.4.1 ERA5-Land. ERA5-Land is a dataset provided by the European Centre for Medium-Range Weather Forecasts (ECMWF). It provides high-resolution climate and global environmental data for the land surface. The dataset is provided in hourly granularity. Air temperature, VPD, precipitation, skin temperature, soil moisture, potential evapotranspiration, shortwave radiation, long-wave radiation are applied as features.

3.4.2 Köppen-Geiger. Based on temperature and precipitation patterns, Köppen-Geiger classifies climates. The original was developed by Wladimir Köppen, followed by the update by Rudolf Geiger. Köppen represents the major climate group by the average annual temperature and precipitation. Köppen-sub is the subcategory of

[1] canopy structure that are measured in remote sensing

[2] downwelling solar radiation

Köppen, and it provides detailed characteristics on seasonality and precipitation.

4 MODELS AND METHODS

This section discusses the general form of the models selected for development in this study to enable GPP upscaling.

4.1 Temporal Fusion Transformer (TFT) Model

The Temporal Fusion Transformer (TFT) is an attention-based time series forecasting model that was first proposed by Google Cloud AI in 2019 [17]. The model can learn short- and long-term temporal patterns by having the encoders to capture the temporal dependencies of input time series data for decoders to generate predicted output sequence. As shown in Figure 3, the model is capable of handling various types of real-world time series data, including static metadata and time-variant known and unknown real and categorical features, and it provides interpretable insights into temporal dynamics. Various applications [17][31] have shown that TFT outperforms existing state-of-the-art models on a range of time-series forecasting tasks.

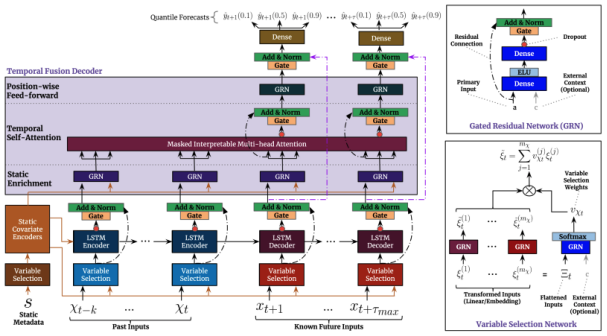


Figure 2: TFT Architecture. [17]

The TFT architecture includes the following novel components and characteristics (Figure 2). 1. Static covariate encoder: The component allows the network to condition temporal forecasts on static metadata by integrating them into three different contexts in the decoder: temporal variable selection, local processing of temporal features (LSTM), and enriching of temporal features with static information in the decoder. 2. Gating components: The components enable skipping over unnecessary parts of the network, reducing the amount of non-linear processing required and resulting in a more efficient model. 3. Variable selection: This process is applied to both time-variant and time-invariant input features to select relevant input features for the prediction problem and to remove unnecessary noisy inputs, resulting in improvement in model performance by utilizing learning capacity only on the most salient features. The variable selection weights also enable variable importance analysis for both encoder and decoder. 4. Temporal processing: The model learns short- and long-term patterns through sequence-to-sequence (LSTM) and multi-head attention-based layers, with additive aggregation of all heads to enhance model explainability through identifying temporal dynamics like prominent seasonality and significant events in the data. 5. Quantile predictions: TFT can produce predictions with a range of likely intervals across all prediction horizons, providing valuable information about the uncertainty of the forecast.

4.1.1 Applying TFT Model. Given I unique entities in a time series data – which corresponds to the flux towers in this study, each entity i (i.e. each flux tower site) is associated with the set of static covariates $\mathbf{s}_i \in \mathbb{R}^{m_s}$, inputs features $\boldsymbol{\chi}_{i,t} \in \mathbb{R}^{m_x}$ and scalar targets $y_{i,t} \in \mathbb{R}$ at each time step t. The input features can be subdivided into two categories: Observed inputs $\boldsymbol{z}_{i,t} \in \mathbb{R}^{m_x}$ which are known before time t but unknown after, while known inputs $\boldsymbol{x}_{i,t} \in \mathbb{R}^{m_x}$ are predetermined(e.g. the hour of the day at time t) and thus remains known after time t. Each quantile forecast can be represented as :

$$\hat{y}_i(q,t,\tau) = f_q(\tau, y_{i,t-k:t}, \boldsymbol{z}_{i,t-k:t}, \boldsymbol{x}_{i,t-k:t+\tau}, \mathbf{s}_i) \quad (1)$$

where $\hat{y}_i(q,t,\tau)$ is the predicted q^{th} sample quantile of the τ-step-ahead forecast at time t and $f_q(.)$ is the TFT prediction model. The model produces forecast for τ_{max} time steps ahead by incorporating past target values and observed values in a finite look-back window with length k (i.e. $y_{i,t-k:t} = \{y_{i,t-k}, ..., y_{i,t}\}$) and known input across the entire time range (i.e. $\boldsymbol{x}_{i,t-k:t+\tau} = \{\boldsymbol{x}_{i,t-k}, ..., \boldsymbol{x}_{i,t}, ..., \boldsymbol{x}_{i,t+\tau}\}$).

In the application of reconstructing the past global GPP, each location (ie flux tower site) would be an entity i. Using Figure 2 as a reference, remote sensing and meteorological data are expected to be time-varying known features because they are known before and after the prediction time. Time information like month, day, hour, of the record are be known at the prediction time so they are also part of the time-varying known variables. Metadata about each location, such IGBP and Köppen labels, are passed into the model as static covariates. Since past global GPP reconstruction modeling focuses on providing an estimate on a single point in the timeline, the decoder length (τ) of the TFT model would be set to 1. On the other hand, encoder length (k) would be a hyperparameter to optimize model complexity and performance.

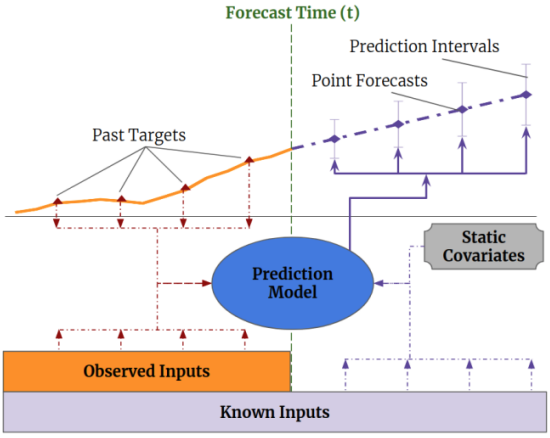


Figure 3: Illustration of a typical application of TFT model: multi-horizon timer series forecasting with static covariates time-dependent past-observed and apriori-known future inputs[17]

4.2 Random Forest Regressor (RFR)

The Random Forest model is a tree-based learning method first introduced by Leo Breimain in 2001 as an extension of the classification and regression tree (CART) algorithm. The random forest model is an ensemble of decision trees where each tree is trained on a random subset of training data with replacement and a random subset of input features. Each tree is trained independently and different trees can learn to perform different tests, resulting in different estimates of the target for the same input. The use of random subsamples is intended to reduce overfitting and improve generalization, making the model well-suited for high-dimensional, noisy datasets. The number of estimators, n_estimators, is an important hyperparameter to tune; a larger number of estimators leads to greater stability, but can also become computationally expensive and performance typically plateaus as the number of estimators increases. Final predictions are obtained by averaging across each individual tree in a process called bagging, which is an acronym for bootstrap aggregating [6].

4.3 XGBoost

The extreme Gradient Boosting method, XGBoost for short, is an efficient and powerful tree-based learning method first introduced by Tianqi Chen and Carlos Guestrin in 2016. XGBoost is designed to optimize large-scale, heterogeneous data by iteratively building an ensemble of weaker decision trees. Compared to the Random Forest method, which uses bagging to reduce overfitting, XGBoost employs boosting. The boosting technique is designed to incrementally improve the model by sequentially adding trees that correct for errors made by previous trees. Overall, the XGBoost model is suitable for a wide range of applications and has gained great popularity for its scalability, interpretability, and high performance [8].

4.4 Evaluation metrics

To assess the performance of the GPP models, a comprehensive evaluation of various metrics was conducted. In alignment with the past literature, three metrics were defined; 1) Root Mean Squared Error (RMSE), 2) Mean Absolute Error (MAE), and 3) Nash-Sutcliffe Efficiency (NSE) were applied to assess the performance of the each model.

4.4.1 Nash-Sutcliffe efficiency (NSE). Nash-Sutcliffe efficiency (NSE) is a normalized statistic that determines the relative magnitude of the residual variance, in comparison with the measured data variance [20]. It is used to evaluate the performance of hydrological and environmental models that are based on simulating observed data. The NSE measures the ratio of variance of the observed data to variance of the simulated data. The metric ranges from negative infinity to 1, where a value of 1 indicates a perfect match between simulated and observed values, and a value of 0 or less indicates that the model is no better than using the observed mean. While the metric has its origin in Hydrology, it is commonly applied in upscaling tasks [14][27][5].

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

4.4.2 RMSE and MAE. Another core metric is the Root Mean Square Error (RMSE), which is calculated as the square root of the mean of the squared differences between the predicted and observed values. RMSE is common in assessing continuous numerical output. Squaring the differences serves to eliminate the influence of negative values, ensuring that all errors are treated as positive values. It also magnifies larger errors. On the contrary, Mean Absolute Error(MAE) is the mean of the absolute values of differences between predicted and observed values.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{4}$$

4.4.3 Quantile Loss. The quantile loss function, $L_\tau(y, \hat{y})$, evaluates the error between predicted values ($\hat{y}_i$) and actual values (y_i) for a specific quantile τ (ranging from 0 to 1). It calculates the average weighted difference over n data points, accounting for the relative positions of predicted and actual values concerning a given quantile. The indicator function, $\mathbb{I}(\cdot)$, serves as a weighting factor, being 1 if $y_i < \hat{y}_i$ and 0 otherwise. One benefit of using quantile loss is the robustness to outliers, and its characteristics of not heavily being influenced by extreme observations. Quantile loss is the commonly applied loss function in the TFT model.

$$L_\tau(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(\tau - \mathbb{I}(y_i < \hat{y}_i))(y_i - \hat{y}_i) \tag{5}$$

5 PREPROCESSING

Due to the time and computational resource constraints, the period of data for modeling was scoped from 2010 to 2015. This is because the period contains better available observations, which is equivalent to having more applicable sites to the modeling. As shown in Figure 1, there is a significant drop on the number of active flux towers after 2015. To limit the impact of gap-filling and to retain the year-long seasonality within the dataset, the flux towers with less than a year of available data and/or more than 20% of missing records in its observation duration are removed, resulting 129 flux tower sites in the dataset.

5.1 Global Time Step

The TFT model in the PyTorch Forecasting[3] library requires all time-series in the dataset to share the same time index. Hence this study referenced the Unix epoch time convention and assigned the number of hours passed since 00:00:00 on January 1st, 1970 as the global time step to each record.

5.2 Imputation and Gap-Filling

A major challenge about the dataset is the considerable amount of missing data, which is likely a consequence of using multiple data

[3]Pytorch Forecasting Temporal Fusion Transformer: https://pytorch-forecasting.readthedocs.io/en/stable/api/pytorch_forecasting.models.temporal_fusion_transformer.TemporalFusionTransformer.html

sources, each of which have varying factors for outages or missing data at a particular time.

Furthermore, there are numerous gaps (i.e. time steps without any records) in the available record sequences per site. These gaps are likely to be caused by flux site shutdowns due to environmental factors, equipment malfunction, maintenance, or power outages. These missing time steps pose a challenge for time-series models, including TFT model, as they expect a uniform sequence of time-series input without a missing time step.

This study looked into a couple imputation/gap-flling methods, including K nearest-neighbors (KNN) and forward-fill, to address the missing data challenges. The forward-fill imputation method fills in missing values or time steps using the most recent record. This approach is effective when the most recent record is within a few time steps of the record to be filled, but can introduce more uncertainty when the time gap between the most recent record and the target record becomes larger. In contrast, the KNN method fills in missing values by considering the available features for a given record, finding a set of neighbors that are most similar along those features (in terms of euclidean distance), and filling in missing values by averaging across the neighbors. This approach is particularly useful because it takes into account the temporal dependencies between data points, and can fill in large blocks of missing values while preserving the continuity and smoothness of the time series. Therefore, this study applied two distinct KNN imputation objects to fill in missing values for existing time step records and gap records separately; both KNN objects used K=5 neighbors, euclidean distance, and a uniform average of neighbor values. The approach was determined based on testing multiple imputation options and evaluating the validation performance of the RFR model. Lastly, the gap-filling flags, indicating whether a record is gap-filled or not, were includes as input feature as the only observed input in all TFT experiments in this study.

5.3 Stratified Train/Test Split

With the limited number of flux towers in the dataset, it is crucial to maintain enough diversity in training, validation, and test datasets so that the model can generalize globally and also be evaluated objectively. Bodesheim implemented a leave-one-site-out cross validation and fit 222 unique models to generate predictions for 222 available sites([5]). This approach bypasses the need to evenly split the dataset into train, validation and test groups; however, hyperparameter tuning is not valid in this approach due to leakage, and it also requires immense compute power to train such a large quantity of distinct models. Building the number of models that corresponds to the number of sites is applicable in RFR, however, required resource becomes challenging to apply the operation to TFT. Given these constraints, the train-validation-test split approach of this study was stratified by a generic IGBP grouping shown in Table 1, leaving 78 sites for training, 26 sites for validation, 25 sites for testing. The locations and IGBP labels of the sites in each dataset is available in Appendix A.

6 EXPERIMENTAL DESIGN

Experimental design was composed of four steps. First, a baseline was built with RFR with hourly preprocessed features. In addition, cross validation was conducted with RFR to observe the potential volatility of model performance by having different combinations of sites through the stratified split. Secondly, feature engineering was implemented with tree models (RFR, XGBoost) to study which features were likely to contribute to the prediction. Impact of engineered features were also studied. Third, TFT models with past GPP values were built as the benchmark. GPP values were applied as both the predictor variable and the taraget variable. This was conducted to observe the best possible performance by TFT, and to investigate the mechanism and dynamics of model behavior, utilizing the strength of interpretability of the TFT. Finally, TFT models for upscaling were built through two different approaches.

6.1 Baseline

The Random Forest Regressor (RFR) algorithm was selected as the baseline for two reasons. Applying the same machine learning algorithm as the past research clarified the impact of having different predictor variables or dataset. Recognizing the difference in the earlier phase helped evaluating different models in the later phase of the study. Another objective was to compare the performance of non-time-aware algorithms (RFR) with time-aware algorithms(TFT) in use of the common dataset. For the hyperparameters, 50 estimators, max tree depth of 10, and the square root of total feature count were used for each individual tree.

Another implementation was the cross validation. Having limitation in available number of groups(Ex. sites) generally makes the modeling and evaluation challenging. Stratified sample split was conducted to generalize the model and to improve the objective evaluation. However, there still existed the concern of having biased model whose performance were highly dependent on train/test split. On the flip side, limitation of resources to run large amount of TFT attempts made it challenging to incorporate the cross validation(CV) to the operation of model development. In order to address this conflict, cross validation was conducted with baseline RFR model to assess the potential impact of reporting specific split group as the final result of the study.

6.2 Feature Engineering with RFR/XGBoost

Experiments for tree models were not limited to model development. Feature engineering was conducted in three perspectives. First agenda was to study the impact of creating additional seasonal features such as hemispheres or seasonal trends. Particularly, past studies reported better performance of their model of mean seasonal cycles(MSC) than each of their models with the higher resolution[27][5]. Since tree algorithms are not assumed to learn time-based transitions and temporal relations among observations, these features were added to expect the potential complement of the general chronological trend that rotationally happens. Second agenda was to find influential features in non-time-aware tree models. Features importances function in both RFR/XGBoost was utilized. Third agenda was to find the best combinations of predictor variables through dimensionality reduction. Utilizing RFR also made a significant contribution to save the computational cost and large amount of time by learning a few cases of behavior in machine learning models. As a subsequent step, findings from the tree models were attempted to TFT through two experiments. One

Category	Covered IGBP	Number of Sites
Evergreen Needleleaf Forests	Evergreen needleleaf forests (ENF)	24
Evergreen Broadleaf Forests	Evergreen broadleaf forests (EBF)	6
Deciduous Broadleaf Forests	Deciduous broadleaf forests (DBF)	20
Mixed Forests	Mixed forests (MF)	7
Shrublands	Closed shrublands (CSH) Open shrublands (OSH)	12
Savanas	Woody savannas (WSA) Savannas (SAV)	12
Grassland	Grasslands (GRA)	24
Cropland	Cropland (CRO)	15
Wetland	Permanent wetlands (WET)	9

Table 1: Generic IGBP grouping for stratified train/validtaion/test split.

No.	Model Name	Algorithms	Description
1	Random Forest Regressor (Baseline)	Random Forest Regressor (RFR)	• Model built with original full features. • Cross validation being conducted to observe the deviation of performance by stratigied split groups.
2	Tuned RFR/XGBoost	Tree-based Regression	• Conducted hyperparameter tuning and feature selection. • The best model was applied as a part of input values in Experiment 5.
3	GPP-TFT	Temporal Fusion Transformer (TFT)	• Ideal TFT application where historical GPP measurements were available. • Included historical GPP measurements as an observed input feature. • Served as a benchmark only as the model was not upscaling-capable.
4	No-GPP-TFT	Temporal Fusion Transformer (TFT)	• Upscaling-capable model. • TFT application where historical GPP measurements are not available. • Excludes historical GPP as an input feature.
5	Tree-FT	Hybrid: RFR + TFT	• Upscaling-capable model. • TFT application for upscaling; where historical GPP measurements are not available. • Includes the *estimated* historical GPP measurements (predicted values from Experiment 2) as an observed input feature.

Table 2: Experiment Summary

was to follow the similar process of dimensionality reduction as tree models, and the other attempt was to apply top features of tree models to TFT models to observe the performance. Having RFR as a core tree model to experiment, XGBoost was studied in parallel to compare model performance and behavior between two tree models.

6.3 GPP-TFT with Historical GPP (Benchmark)

The initial TFT modeling approach followed the default definition of the TFT model (Equation 1) under the ideal scenario in which the past measurements of GPP were available at the time of the prediction. While the final goal of this research was upscaling, which equaled the condition that past measurements of GPP of the target sites were unavailable, we built the best possible benchmark achievable by TFT to study the mechanism and dynamics of model behavior; influential features by time, temporal transition of attentions and key hyperparameters such as as encoder length. The goal of this experiment was to understand the maximum potential performance of TFT model by having past targets as model inputs. In the later sections, the scope of the study shifted to upscaling, which corresponded to building a TFT model without past GPP targets as predictor variables. Instead, the TFT model was built by remote sensing and meteorological features to predict GPP at each timestep.

In the original study of TFT [17], most experiments used around 500K of records for training and all the hourly forecast experiments used 168 (1-week) as the encoder length. Therefore, The initial GPP-TFT experiment subsetted the first year-long of records from each flux tower site, resulting in around 683K records, to ensure that the model has visibility to data with year-long seasonality while remaining feasible to train. After instantiating a model with this configuration, tuning was applied to key hyperparameters like hidden layer size and dropout rate. Subsequent experiments on various encoder lengths, 2 weeks, 3-day and 1-day, was conducted using the same set of hyperparameters to examine the effect of encoder length on model performance and interpretation. Concurrently, significant features in both encoder and decoder, temporal transition of attention by features, and overall trend of attention scores by different model setups were analyzed to investigate the model behavior and potential factors behind the model performance.

6.4 Upscaling Models - Tree-FT & No-GPP-TFT

In real-world application of GPP upscaling, historical measurements of GPP is not available for locations with no flux tower. Therefore, the study proposed two unconventional applications of TFT to overcome the challenge of missing historical target value; 1) No-GPP-TFT Model: Completely remove past target values from TFT model input, 2) TREE-TFT Model: Use the ideal/initial TFT model trained with past GPP as input and substitute the entirely blank past target value by estimated historical GPP of either RFR or XGBoost. Using model performance and interpretation resulted from the ideal scenario in the first experiment, both experiments went through an iterative hyperparameter tuning process. Finally, feature engineering; particularly dimensionality reduction was implemented to study whether the model improvement was expected.

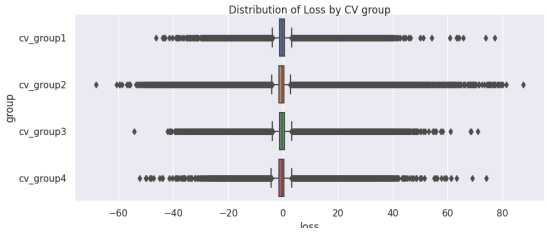


Figure 4: Distribution of Loss by CV group

6.5 Interpret TFT Models

As is discussed in the earlier section, one of the strengths of TFT is the capability of the model interpretability. While feature importance is available in tree models, TFT can also detect the influence of each feature in temporal resolution. Furthermore, attention is available in temporal resolution, which enables scientists to analyze the most impactful time point of the encoder. Since the conventional machine learning algorithms have always argued the black-box structure of the decision making process, TFT has a capability to provide better transparency of the mechanism of a model. There are three topics studied in the analysis section.

7 RESULTS

7.1 Cross Validation Attempt in RFR

Since stratified split divided the Flux Tower sites to five groups in the the earlier phase, 4-fold split CV were conducted by using four of the groups to CV[4]. One of the CV group (CV Group4) was originally assigned as a group that being applied to every model[5]. Model performance of each CV group(Table 3) as well as their distributions of loss (Figure 4) show negligible differences among the CV groups. This finding led to persist the condition of reporting model performance of CV Group 4 as the model result.

The model performance metrics (Table 3) and loss distribution of each CV group (Figure 4) suggest the variances among them are negligible. This finding, along with the compute and time constraints of this study, led to the conclusion that cross-validation with TFT is not necessary.

7.2 Model Comparison

Results of the models are shown in the Table 4. Model 1 refers to the baseline of the study with RFR, and Model 2 refers to the XGBoost. Both were trained on the full set of original remote sensing and meteorological features. Model 3 was built with engineered MSC features as well as a hemisphere indicator, beside the original full features. Dimensionality reduction was implemented in Models 4 and 5, and each of optimal number of features were applied to the modeling. Model 6 refers to the benchmark GPP-TFT model, which appiled past GPP values as predictor variables. Models 7 and 8 refers to Tree-FT models, and either the predicted values from RFR or XGBoost was applied to TFT as a part of the input features. Model 9 to 14 refer to an another upscaling model that has no past

[4]The rest of one group was left as the final test set.
[5]After the CV implementation, CV Group4 was consistently used throughout this study

CV Group	RMSE	MAE	NSE
CV Group1	3.657	1.979	0.681
CV Group2	3.644	1.963	0.683
CV Group3	3.650	1.975	0.682
CV Group4	3.675	2.011	0.678

Table 3: CV Result of RFR Baseline(Test Result)

GPP values used as input features. The difference among 9 to 14 was the key hyperparameters.

Several models developed in this research resulted in better performance metrics than those from the past literature[5]. While the best performing model was the Model 4, RFR model fit on Top 9 features, both the Tree-FT and No-GPP-TFT models also improved upon the benchmark of previous work. The details of these individual models and results are discussed below. The significant difference from the past literature was the expanded feature set for upscaling. While finally the best model in RFR converged to 9 features, it was obtained as a result of optimization with large amount of options. Tuning the model hyperparameters also contributed to marginal improvements.

7.3 Feature Engineering with RFR/XGBoost

7.3.1 Impact of engineered features. A second variant of the RFR model was trained on a dataset with the original feature set as well as a set of mean seasonal cycle (MSC) features for the following remote sensing predictors; temperature(T-ERA), shortwave radiation(SW-IN-ERA), precipitation(P-ERA), EVI, NDVI, and NIRv. By adding them to original features, they were expected to capture the historical seasonal patterns in a given site. While the past literature worked on modeling MSC([5], [27]), this study add the information as supplemental features of RFR. The amplitude of MSC features was also utilized to capture the range in seasonal cycles for each site. Additionally, a hemisphere feature was engineered from the latitude variable. Since the seasonal trend by northern/southern hemisphere was observed in both GPP as well as loss, the feature was add in to the predictors. However, as is described in Table 1, having every additional features in the model did not improve the baseline (Model 3).

7.3.2 Influential features and dimensionality reduction. Using the built-in functions of RFR and XGBoost, feature importances were examined in both tree-based models (Appendix Figure 20, Figure 21b). A common trend of both tree models was the limited number of influential features. Furthermore, SW-IN-ERA (Shortwave radiation), NDVI, NIRv, Lai(Leaf Area Index), EVI(Enhanced Vegetation Index) were common top ranked features in the both models. Differences between the two tree models was the significance of hour, TA-ERA(Air Temperature) or VPD(Vapor Pressure) in RFR, and having MODIS IGBP in XGBoost. Since the most influential features were limited to top ranked variables, dimensionality reduction was experimented. In RFR, having top 9 features (SW-IN-ERA, NDVI, NIRv, hour, Lai, TA-ERA, VPD-ERA, EVI, CSIF-SIFdaily(Daily Average Solar-Induced Fluorescence)) improved the test NSE in 0.704 (Table Figure 4, Model 5). More surprisingly, only having to top 3 features (NDVI, NIRv, Shortwave radiation) led to obtain best model in the XGBoost. Its test NSE was 0.677 (Table Figure 4, Model 6). These result implies the number of influential features are limited in tree models. When the correlations among features were examined during the initial phase of the study, many of the features were correlated(Appendix: 4). This implies the potential colinearity of correlated features, which may led the improvement by dropping those features. Interestingly, the correlation varied by IGBP type. Observing that some features have correlation in IGBP type and the others are not, significant features may vary by land cover types. Features importance is discussed further in the later section.

7.4 GPP-TFT Model Analysis

The Benchmark GPP-TFT model, which was fit on data that includes historical GPP as model input achieved significantly stronger performance than any preceding benchmark. As expected, this strong performance goes to show that present GPP is more easily predicted by historical GPP values. Unfortunately, this model variant is not useful for the upscaling task as the historical GPP values will not be available globally across time until the entire globe is cover by flux tower sites. Despite this, this exercise still proved valuable as it helped to set quality expectations for the other TFT models trained without past GPP input. It helps to understand what the maximum model performance would be with near-perfect information, which in turn sets the relative scale for measuring improvements between current state and ideal state. Additionally, the interpretable outputs of this model help to gain an understanding of what trends should be expected of the upscaling models.

7.5 Tree-TFT Model Analysis

Given the strong performance of the Benchmark GPP-TFT model, which was trained with historical GPP as input, it was hypothesized that the model might show the better performance by inputting estimated historical GPP values, compared to model that has none of the clue about historical GPP values as input. The goal of this Tree-FT experiment was to maximize performance, while qualifying a model for upscaling to global regions where historical GPP inputs were unavailable. In this experiment, the tuned RFR and XGB models were each evaluated for use in generating the input GPP predictions for the TFT model. Interestingly, the Tree-FT model variant that used the RFR predictions for past GPP inputs had much stronger performance than the Tree-FT variant that used XGB predictions as input. In short, this experiment did not achieve significantly different performance from the No-GPP-TFT or tree-based baseline models. One potential explanation for this could be that despite the capabilities of the TFT architecture, the global dataset consists of many diverse regions and climates that are very

No.	Model	Features	Train	Hidden Size	Encoder	Decoder	RMSE	MAE	NSE
–	**Previous Work**								
0	Bodesheim et al. 2018 (Upscaling RFR)	-	-	-	-	-	3.940	-	0.670
–	**Baseline**								
1	RFR-BASELINE	Original	6 year	-	-	-	3.675	2.011	0.678
2	XGB-ORG	Original	6 year	-	-	-	3.532	1.855	0.690
3	RFR-ENG	Original + Engineered	6 year	-	-	-	3.752	2.111	0.664
4	RFR-TOP7	Top 9 features*	6 year	-	-	-	3.523	1.841	0.704
5	XGB-TOP3	Top 3 features**	6 year	-	-	-	3.610	1.870	0.677
–	**Benchmark TFT with Past GPP**								
6	GPP-TFT-14E5T	Original	1 year	136	24*14	1	2.132	1.016	0.886
–	**Upscaling Tree-FT**								
7	RFR-TFT-14D	Slim Features***	5 year	16	24*14	1	3.630	1.900	0.671
8	XGB-TFT-14D	Slim Features	5 year	16	24*14	1	3.807	2.002	0.638
–	**Upscaling No-GPP-TFT**								
9	No-GPP-TFT-3D-16HS	Slim Features	5 year	16	24*3	1	3.618	1.906	0.673
10	No-GPP-TFT-7D-16HS	Slim Features	5 year	16	24*7	1	3.594	1.904	0.677
11	No-GPP-TFT-7D-64HS	Slim Features	5 year	64	24*7	1	3.618	1.917	0.673
12	No-GPP-TFT-7D-16HS-FULL	Original	5 year	16	24*7	1	3.682	1.880	0.662
13	No-GPP-TFT-14D-16HS	Slim Features	5 year	16	24*14	1	3.691	1.936	0.660
14	No-GPP-TFT-30D-16HS	Slim Features	5 year	16	24*30	1	3.912	2.089	0.609

Table 4: Comparison of test set performance metrics of RFR, XGBoost and TFT models
***TOP9: SW-IN-ERA, NDVI, NIRv, hour, Lai, TA-ERA, VPD-ERA, EVI, CSIF-SIFdaily**
****TOP3: NDVI, NIRv, SW-IN-ERA**
*****Slim Features: TA-ERA, SW-IN-ERA, LW-IN-ERA, VPD-ERA, P-ERA, PA-ERA, NDVI, b2, b4, b6, b7, BESS-PARdiff, CSIF-SIFdaily, ESACCI-smi, Percent-Snow, LAI, LST-Day, LST-Night**

difficult to generalize. In the future, if advancements are achieved on the tree-based models, it would be worth re-evaluating the Tree-FT approach to test if higher quality GPP prediction inputs could boost the model performance.

7.6 No-GPP-TFT Model Analysis

The No-GPP-TFT model, in which historical GPP is held out from encoder input, also improved on the state of the art metrics ([5]). For this model, the key hyperparameters were encoder length (3 days, 7 days, 14 days, 30 days), the hidden layer size (16, 64) and learning rate (1e-5, 1e-3). Ultimately, the best performing No-GPP-TFT variant used an encoder length of 7 days, a hidden layer size of 14, and learning rate of 1e-3.

The initial hypothesis for this model stated that it might improve GPP predictive performance by gaining information on weather and climate events for the region in the preceding days. The attention plots shown in Figure 7b indicate that the model learned to identify diurnal cycles, and the loss plots shown in Figure 5 also show that the model converged without overfitting.

While the No-GPP-TFT model performed better than previous benchmark, it did not significantly differentiate from the tree-based

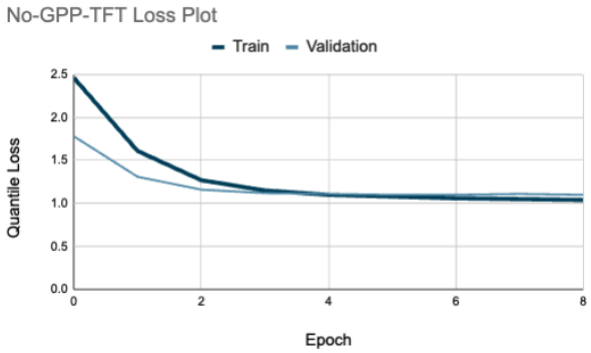


Figure 5: Learning curve of No-GPP-TFT Model with 7-day encoder length.

models explored. One potential factor for this lack of differentiation from tree-based methods could be that the dataset is global and very difficult to generalize effectively. Even though the model learned to identify diurnal cycles effectively (as shown in Figure 7),

it might have struggled to extract the relevant predictor features from individual sites that vary significantly across the globe.

8 TFT MODEL INTERPRETABILITY

As is described in the earlier section, one of the strength of the TFT model is the interpretability. The TFT model has capabilities to provide quantified impact of each predictor variable in each prediction. In this section, influential factors in time frame or by predictor variables are analyzed.

8.1 Analysis of Average Attention by Model

A key hyperparameter to tune is assumed for the TFT model is assumed to be the encoder length. This may be assumed as especially relevant in the study of GPP, since studies report the memory function of plants with extreme historical events [18]. Assuming the domain knowledge, GPP models are expected to effectively balance the trade-off between longer-term memory and compute constraints of larger model encoders.

For the benchmark GPP-TFT model, Figure 6 presents the average attention across the historical time steps in encoder. Time 0 is the predicted time[6], and the hourly look-back window adds -1 in each hour until the end of encoder(encoder length is defined as 24*7=168 hours). Since the encoder/decoder slide throughout the data period, values in the x-axis correspond to the relative time index. The plot shows a peak attention spike at 24 hours before the prediction time, and each preceding day has attention peaks of decreasing magnitude across the encoder. Therefore, when historical GPP is used in the model input, the GPP prediction at inference time is most highly impacted by the GPP value from the same hour of a day before.

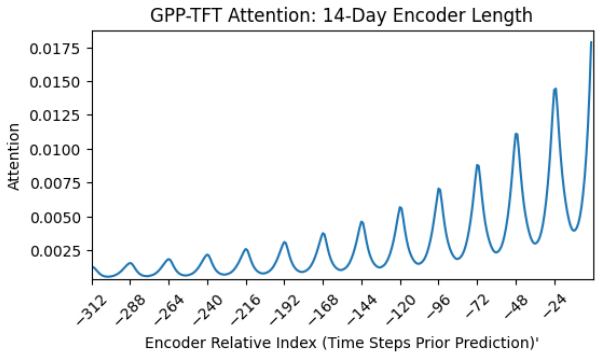


Figure 6: Average attention plot of the benchmark GPP-TFT Model with 14-day encoder length.

As shown in Figure 7a, the No-GPP-TFT model with 3-day encoder failed to learn any diurnal cycles, and had increasing attention for more distant encoder indices. As the encoder length increased to 7 days (Figure 7b), the attention curve had a consistent 24-hour cycle that slowly decreased for each additional day of history [7]. It suggests that diurnal trend was learned by the attention mechanism. On the contrary, the peaks of daily attention were not exactly 24 hours before the prediction time, as seen in the GPP-TFT. Instead, it had peak around 20 hours in advance of prediction. This implies that the model had potentially focused on different factors, such as features, at the given hour than the same hour of the previous day. In Figure 7c, the model with 14-day encoder length follows the same pattern as the 7-day encoder model with clear diurnal cycles decreasing each successive day. When increasing the encoder length to 30 days (Figure 7d), the daily attention cycle was also apparent, and decreases over the first 3 weeks of history. Finally, it began to increase again across the fourth week of the history.

8.2 Model Behavior by IGBP

An ideal GPP inference model would be highly generalized and accurate across global regions. However, generalization is very difficult to attain across many unique regions of Earth that have distinct ecosystems and vegetation patterns. One benefit of attempting to train such a global model is that different trends can be analyzed within the IGBP sub-groups of the dataset.

Figure 8 shows the attention plots of a select few IGBP groups that have unique properties: Woody Savannas (WSA), Open Shrublands (OSH), Permanent Wetlands (WET).

Compared to the OSH group, which has a relatively flat attention curve across encoded timesteps, the WSA (Woody Savannah) and WET (Wetlands) groups have clear daily variations. However, each site has different peaks of daily attention; OSH and WSA (Woody Savanna) has the highest attention nearly 24 hours previous, whereas WET has peak attention 18 hours in advance. It implies the potential unique trends by land cover types. Historical information is processed differently in inference.

Additionally, Figure 8 shows that daily-attention can change over time in different directions. The peak daily attention for WET increased over time while peak daily attention for WSA decreased consistently. This shows that different climate regions operate on different timescales, and each may require a different encoder length to capture the relevant slice of history for inference.

8.3 Influential features

The TFT model provides feature importance by three types; Encoder Variables Importance, Decoder Variable Importance, and Static Variable Importance. This research focused on encoder variables importance for two reasons. First, the decoder length was consistently applied as 1 throughout the study. It led to obtaining decoder attention zero. Secondly, there were only three static variables applied in the study, while having various numeric features as input. PyTorch forecasting API provides the summary of encoder variable importance, howerver, this study attempted to break down the importance by taking snapshot of prediction time t in site-level. Since encoder variables are a part of the input to obtain encoder attention, it was assumed as the inducing factor of the shape of encoder attention. By breaking down the feature importance to the lowest granularity, importance of each feature was able to be plotted across relative time indices of the encoder length. This section consists of 1) Walk through of encoder feature importance with GPP-TFT, 2) Comparison of No-GPP-TFT model and GPP-TFT model, 3) Comparison of GPP-TFT model by IGBP type, and 4) Investigation of seasonality in

[6]This is equivalent to decoder
[7]reading right to left

IGBP	RMSE	MAE	NSE
Evergreen Needleleaf Forests (ENF)	4.025	2.211	0.634
Evergreen Broadleaf Forests (EBF)	4.391	2.607	0.050
Deciduous broadleaf forests (DBF)	3.315	1.742	0.859
Mixed Forests (MF)	3.441	2.154	0.650
Shrublands (Open Shrublands (OSH))	1.252	0.698	0.193
Savanas (Woody savannas (WSA))	2.933	1.632	0.595
Grasslands (GRA)	2.939	1.525	0.733
Cropland (CRO)	5.246	2.941	0.536
Wetlands (WET)	3.900	2.126	0.571

Table 5: No-GPP-TFT model evaluation metrics breakdown by IGBP types.

feature importance, 5) Differences between daytime and nighttime prediction.

8.3.1 Walk through of encoder feature importance with GPP-TFT. An example of feature importance in GPP-TFT model(Figure 9) shows a sample of encoder variables applied to predict the specific time point in AU-DaP(Daly River Savanna, Australia). GPP-TFT with full features, and 14 days of encoder length was applied there. The figure includes two pieces of information. The bold line plot corresponds to an encoder attention in AU-DaP, and the other stack area plot shows the stack of feature importance of the top 15 features. Legend of the stack plot aligns the order of the stack area plot. Higher rank of the importance is located at the bottom of legend, therefore gap-flag-hour is equivalent to the most influential feature of this snapshot[8]. For the prediction time t, the first record of 220K[9] records, 0AM of January 15 in 2010 was chosen for the clarity. -24 in the x-axis indicates the attention and the breakdown of importance in 0AM of January 14, 2010.

In regards to each feature, rotational constant spike was observed in b4(MODIS spectral bands). Their peaks were observed in the daytime. One potential factor of gaining the gap-flag-hour as the most influential feature was the impact of missing values in the dataset in AU-DaP. It contained missing values in a specific time range(5AM to 9AM), which may be inevitable for models to learn temporal pattern through the gap-filled records.

8.3.2 Investigation of seasonality in feature importance. Having the common site AU-DaP with different prediction time t, seasonal difference was observed. (Figure **??**). 0AM of August 8, 2010 was chosen to compare the shapes between summer and winter. Since Australia is located in the southern hemisphere, Figure **??** is equivalent to winter, and the former Figure 9 corresponds to summer. As a result, temporal dependent variables such as relative time index, gap-flag-hour, and the past GPP(GPP-NT-VUT-REF) or air temperature(TA-ERA) kept showing the significant presence in both winter and summer, while the ratio of the b4 became much smaller in winter. Accordingly, the daily constant trend became vague, compared to the plot in summer.

Since the AU-DaP is classified as GRA (Grasslands) in the IGBP type, another location with GRA was compared to study whether the seasonal difference was potentially generalized or limited to specific locations (Figure **??**). Another site from validation set with GRA was US-AR1 (ARM USDA UNL OSU Woodward Switchgrass 1, Oklahoma, the United of States).

Besides the difference in season between southern and northern hemisphere, common characteristics were observed across two sites (AUDaP: Figure 9 and Figure 10, US-AR1: Figure 11 and Figure 12). For example, b4 was recorded as the most influential feature in summer in US-AR1 as well. Furthermore, huge spikes were observed during the daytime. On the other hand, the top 7 features, which satisfied more than 50% of the importance, were common features between AU-DaP and US-AR1. To summarize, Grassland sites may have common patterns by season.

8.3.3 Comparison of GPP-TFT model by IGBP type. Up to this point, comparisons were made by different sites with the common IGBP type GRA. In order to compare potential distinction by IGBP type, 1) Deciduous Broadleaf Forests(DBF): US-BaR(Bartlett Experimental Forest, New Hampshire, the United States), 2) Evergreen Needleleaf Forests(ENF): FI-Hyy(Hyytiala, Finland), 3) Croplands(CRO): FR-Aur(Aurade, France) were plotted in Figure 13. August of US-BaR and FI-Hyy show daily spikes of b4. As discussed previously, this trend was observed in US-AR1(August) and AU-DaP(January) as well. On the other hand, b4 in FR-Aur was not as influential as the four earlier sites. Finally, winter feature importance in US-BaR was compared with that of US-AR1, followed by the August feature importance in AU-DaP(Figure 10, Figure 11, Figure 14). Adding to temporal dependent variables, precipitation and air temperature were seen as the potential top features across three IGBP types.

8.3.4 Comparison of No-GPP-TFT model and GPP-TFT model. The GPP-TFT model was also compared with the feature importance of the No-GPP-TFT model(Figure 15, Figure 16). As is discussed in the previous section, the shape of attentions were different between two models. Line up of top features in No-GPP-TFT model had huge difference from GPP-TFT model. Having different lineup in top features, No-GPP-TFT may have a different dynamics of recognizing patterns.

8.3.5 Differences between daytime and nighttime prediction. Finally, difference between daytime and nighttime prediction was compared(Figure 17 and Figure **??**. 12PM was chosen instead of 8AM. While the shape of features did not have difference, b4 is observed

[8]It was followed by b4 and relative-time-index
[9]Total across all the sites. AuDaP was the first site in the validation set

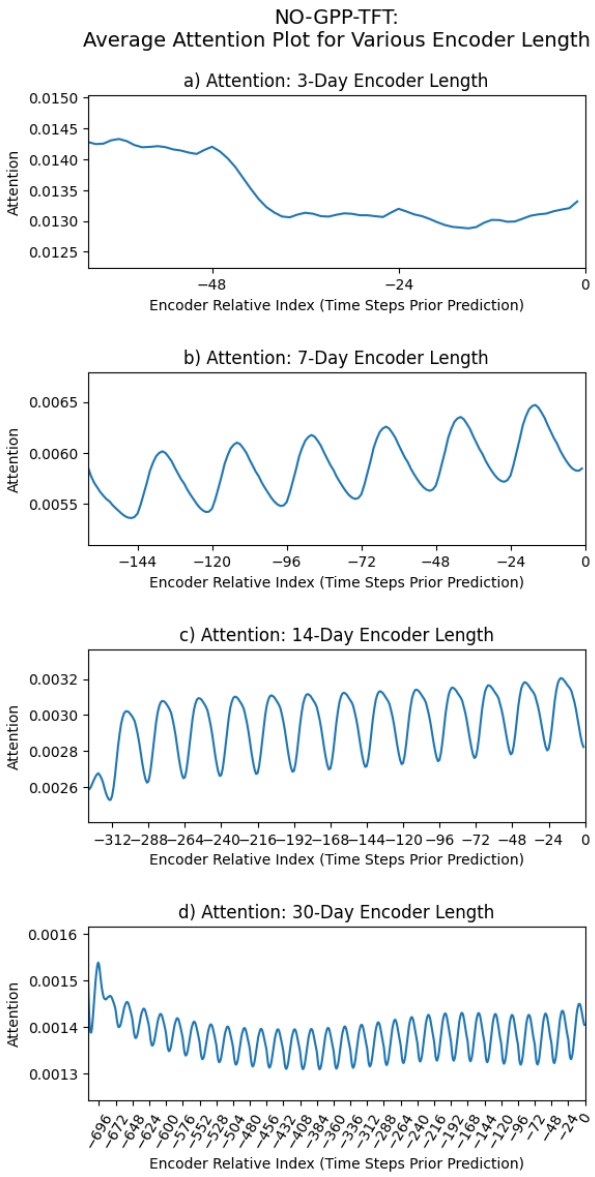


Figure 7: Average attention plot for No-GPP-TFT model of 3-day, 7-day, 14-day, and 30-day encoder lengths.

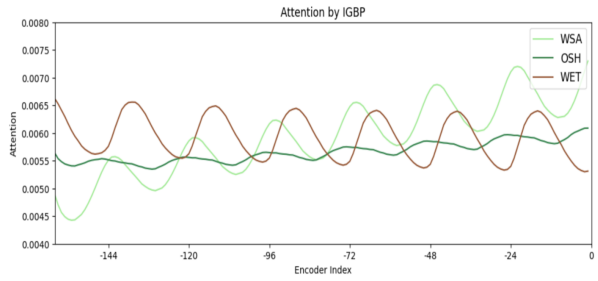


Figure 8: Attention by IGBP (subset)

as the potential significant determinant of the attention in some of the time points.

9 CONCLUSION

This study presented a set of predictive models to perform hourly GPP estimation on global scale for unseen locations. We have introduced a set of time-aware modeling approaches using TFT to estimate hourly values from predictor variables at various temporal and spatial resolution. Our modeling approaches included setting a benchmark of TFT performance by training a model with past GPP values as predictors, as well as training two upscaling TFT models; one trained without past GPP as input, and another two-stage model that inputs GPP predictions from tree-based models into a final TFT model. Compute constraints represented a key limitation of this research, primarily affecting our ability to perform wide hyperparameter tuning jobs, and also cross validation of deep learning models.

When comparing to the benchmark performance established in previous work, several of our tree-based baseline models and TFT-based upscaling models resulted in improved performance. Despite marginal performance improvements, the upscaling TFT models did not generalize well across the globe and were not able to surpass efficiency of 0.7. Such result perhaps is indicative of the limitation with a single-model solution to a complex problem such as global GPP upscaling. However, this finding along with analysis presented in this study indicate GPP estimation could be further improved by training distinct models for each unique IGBP group. This study provided the foundation for future solutions using TFT and the new form of interpretation based on TFT model's feature importance and attention outputs that are applicable not just to carbon flux research, but also other upscaling scenarios in other industries, where past target values do not presented in the time-series dataset.

ACKNOWLEDGMENTS

We are grateful to have the guidance and support from to advisors from UC Berkeley School of Information and the Quantitative Ecosystem Dynamics Lab at UC Berkeley. We would like to express our deepest appreciation to Dr. Alberto Todeschini and Dr. Puya H. Vahabi, professors of the Data Science Master's program at University of California, Berkeley. We could not have undertaken this journey without the advice from Dr. Yanghui Kang and Dr. Maoya Bassiouni from the Quantitative Ecosystem Dynamics Lab in the UC Berkeley & Lawrence Berkeley National Lab. We also wish to thank to Alex Carite for assisting with XGBoost modeling.

REFERENCES

[1] Marc Aubinet, Timo Vesala, and Dario Papale. 2012. Eddy Covariance: A Practical Guide to Measurement and Data Analysis. *Eddy Covariance* (2012).

[2] Dennis D. Baldocchi, Bruce B. Hincks, and Tilden P. Meyers. 1988. Measuring Biosphere-Atmosphere Exchanges of Biologically Related Gases with Micrometeorological Methods. *Ecology* 69, 5 (1988), 1331–1340. https://doi.org/10.2307/1941631 arXiv:https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.2307/1941631

[3] Christian Beer, Markus Reichstein, Enrico Tomelleri, Philippe Ciais, Martin Jung, Nuno Carvalhais, Christian Rödenbeck, M. Altaf Arain, Dennis Baldocchi, Gordon B. Bonan, Alberte Bondeau, Alessandro Cescatti, Gitta Lasslop, Anders Lindroth, Mark Lomas, Sebastiaan Luyssaert, Hank Margolis, Keith W. Oleson, Olivier Roupsard, Elmar Veenendaal, Nicolas Viovy, Christopher Williams, F. Ian Woodward, and Dario Papale. 2010. Terrestrial Gross

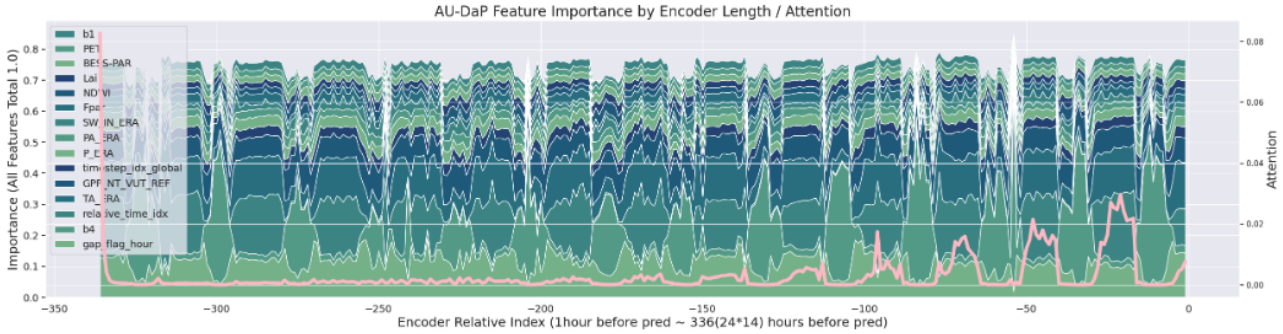


Figure 9: Feature importance with attention GPP-TFT model with 14 days encoder, full features Site: AU-DaP, prediction time t: 0AM of January 15th, 2010

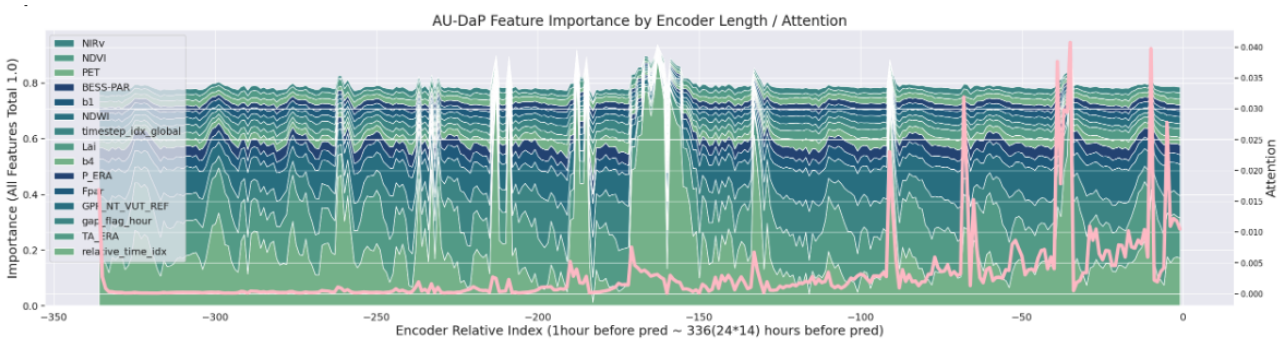


Figure 10: Feature importance with attention GPP-TFT model with 14 days encoder, full features Site: AU-DaP, prediction time t: 0AM of August 8th, 2010

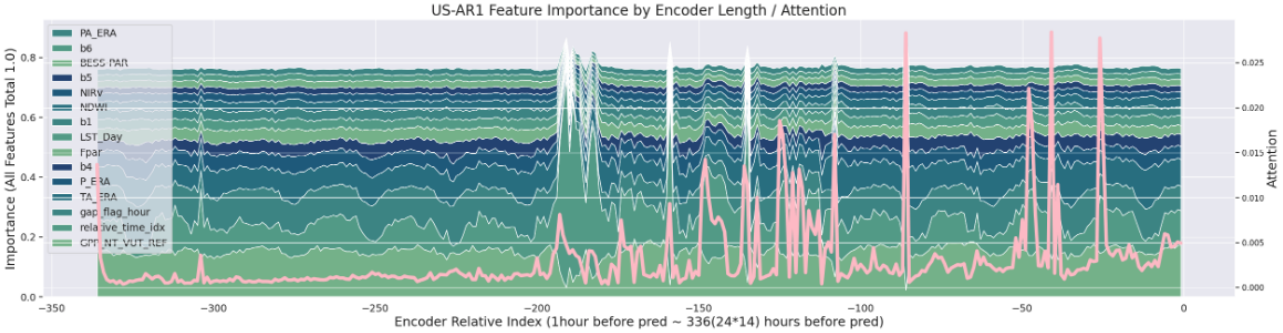


Figure 11: Feature importance with attention GPP-TFT model with 14 days encoder, full features Site: US-AR1, prediction time t: 0AM of January 15th, 2010

Carbon Dioxide Uptake: Global Distribution and Covariation with Climate. *Science* 329, 5993 (2010), 834–838. https://doi.org/10.1126/science.1184984 arXiv:https://www.science.org/doi/pdf/10.1126/science.1184984

[4] Simon Besnard, Nuno Carvalhais, M. Altaf Arain, Andrew Black, Benjamin Brede, Nina Buchmann, Jiquan Chen, Jan G. P. W Clevers, Loïc P. Dutrieux, Fabian Gans, Martin Herold, Martin Jung, Yoshiko Kosugi, Alexander Knohl, Beverly E. Law, Eugénie Paul-Limoges, Annalea Lohila, Lutz Merbold, Olivier Roupsard, Riccardo Valentini, Sebastian Wolf, Xudong Zhang, and Markus Reichstein. 2019. Memory effects of climate and vegetation affecting net ecosystem CO2 fluxes in global forests. *PLOS ONE* 14, 2 (02 2019), 1–22. https://doi.org/10.1371/journal.pone.0211510

[5] P. Bodesheim, M. Jung, F. Gans, M. D. Mahecha, and M. Reichstein. 2018. Upscaled diurnal cycles of land–atmosphere fluxes: a new global half-hourly data product. *Earth System Science Data* 10, 3 (2018), 1327–1365. https://doi.org/10.5194/essd-10-1327-2018

[6] Leo Breiman. 1996. Bagging Predictors. *Mach. Learn.* 23 (August 1996), 123–140. https://doi.org/10.1007/BF00058655

[7] Bo cai Gao. 1996. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment* 58, 3 (1996), 257–266. https://doi.org/10.1016/S0034-4257(96)00067-3

[8] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery,

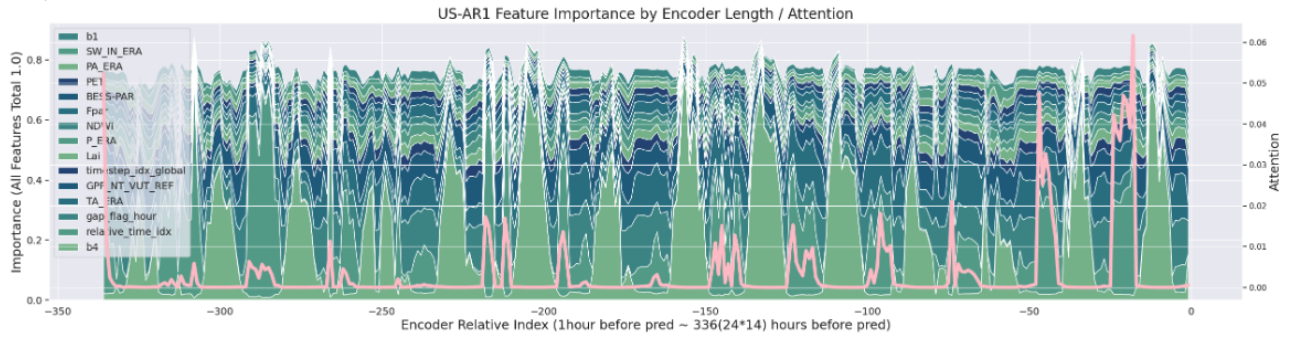


Figure 12: Feature importance with attention GPP-TFT model with 14 days encoder, full features
Site: US-AR1, prediction time t: 0AM of August 8th, 2010

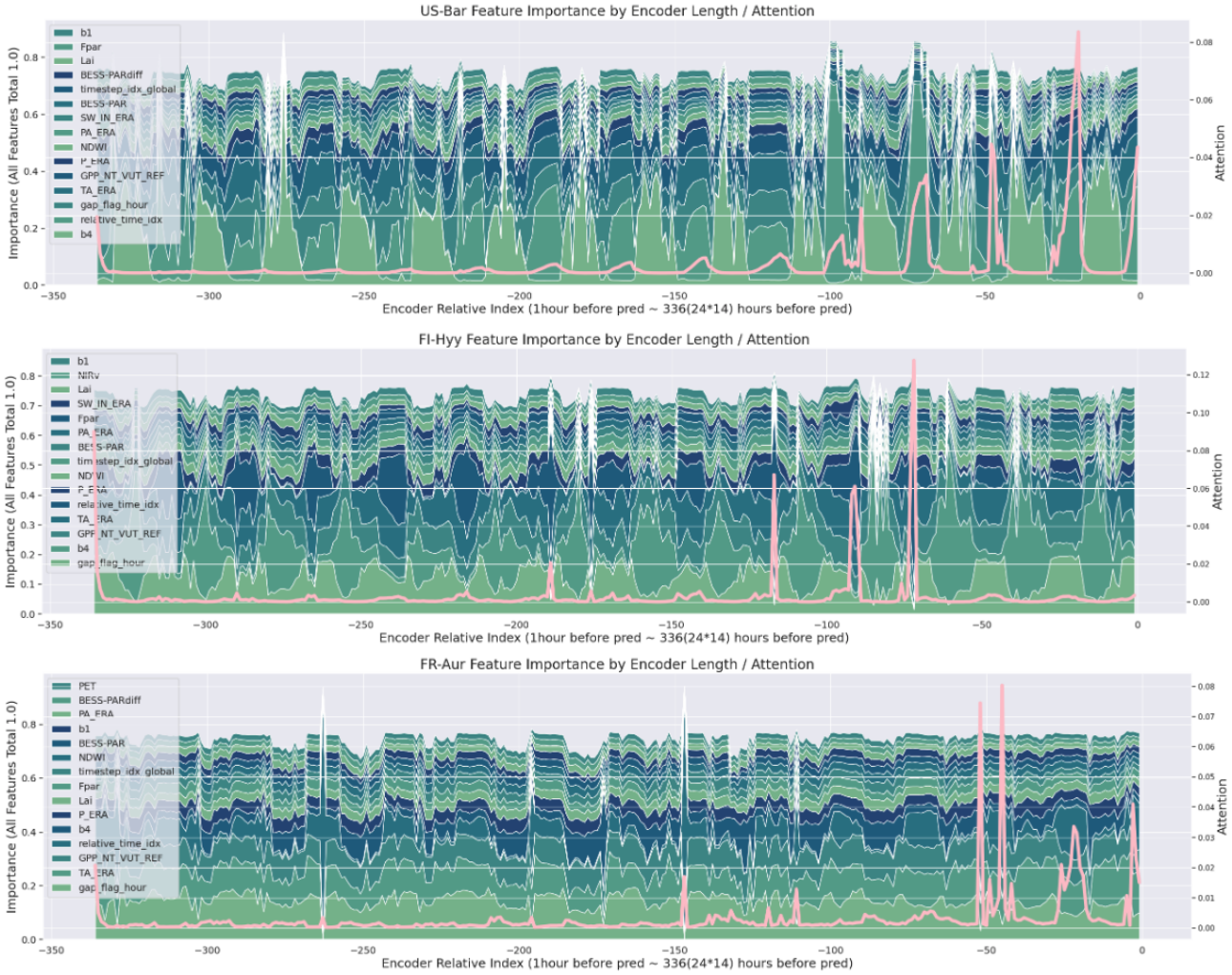


Figure 13: Comparison of Feature Importances by IGBP type
Sites: US-BaR, FI-Hyy, FR-Aur, prediction time t: 0AM of August 8th

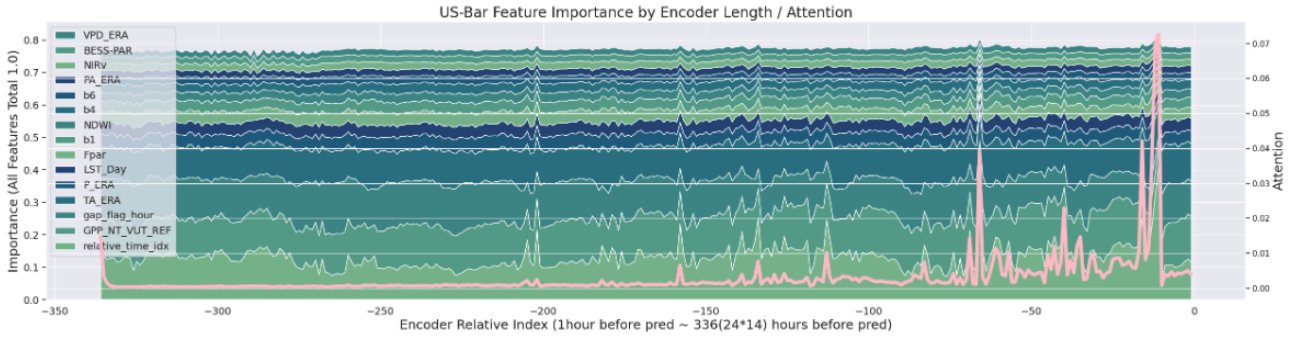


Figure 14: Feature importance with attention GPP-TFT model with 14 days encoder, full features Site: US-Bar, prediction time t: 0AM of January 15th, 2010

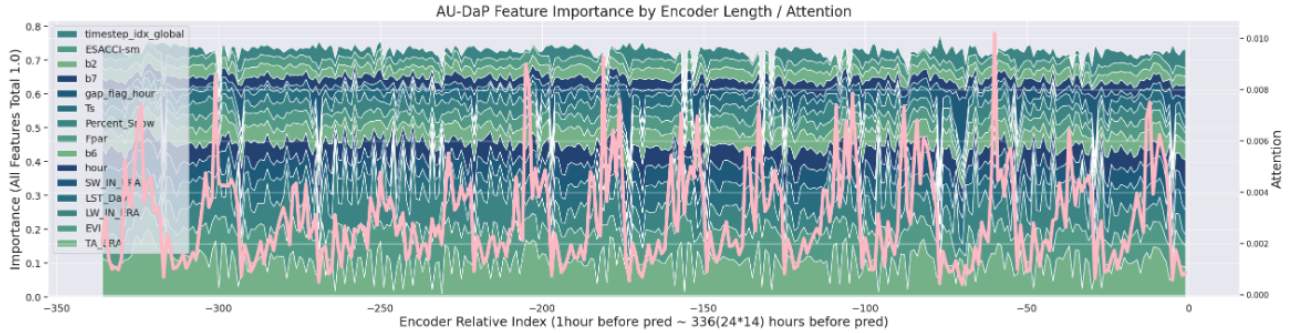


Figure 15: Feature importance with attention No-GPP-TFT model with 14 days encoder, full features Site: AU-DaP, prediction time t: 0AM of January 15th, 2010

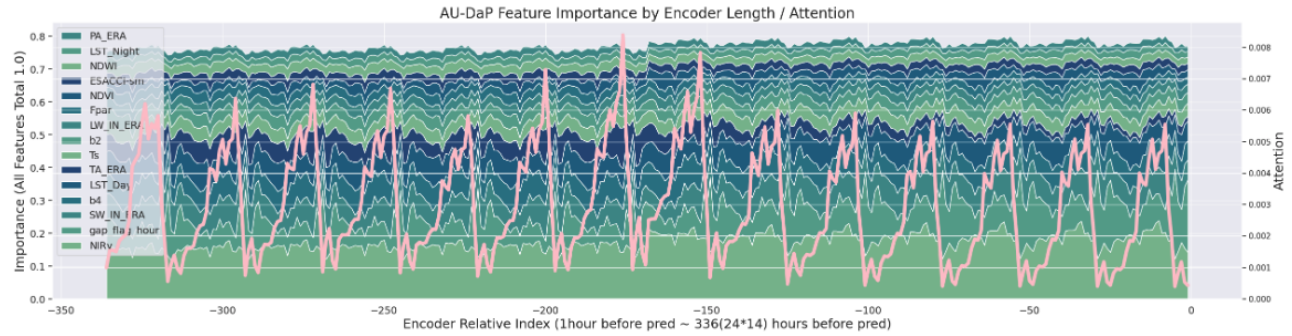


Figure 16: Feature importance with attention No-GPP-TFT model with 14 days encoder, full features Site: AU-DaP, prediction time t: 0AM of August 8th, 2010

New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[9] A Huete, K Didan, T Miura, E.P Rodriguez, X Gao, and L.G Ferreira. 2002. Overview of the radiometric and biophysical performance of the MODIS vegetation indices. *Remote Sensing of Environment* 83, 1 (2002), 195–213. https://doi.org/10.1016/S0034-4257(02)00096-2 The Moderate Resolution Imaging Spectroradiometer (MODIS): a new generation of Land Surface Monitoring.

[10] Kazuhito Ichii, Masahito Ueyama, Masayuki Kondo, Nobuko Saigusa, Joon Kim, Ma. Carmelita Alberto, Jonas Ardö, Eugénie S. Euskirchen, Minseok Kang, Takashi Hirano, Joanna Joiner, Hideki Kobayashi, Luca Belelli Marchesini, Lutz Merbold, Akira Miyata, Taku M. Saitoh, Kentaro Takagi, Andrej Varlagin, M. Syndonia Bret-Harte, Kenzo Kitamura, Yoshiko Kosugi, Ayumi Kotani, Kireet Kumar, Sheng-Gong Li, Takashi Machimura, Yojiro Matsuura, Yasuko Mizoguchi, Takeshi Ohta, Sandipan Mukherjee, Yuji Yanagi, Yukio Yasuda, Yiping Zhang, and Fenghua Zhao. 2017. New data-driven estimation of terrestrial CO2 fluxes in Asia using a standardized database of eddy covariance measurements, remote sensing data, and support vector regression. *Journal of Geophysical Research: Biogeosciences* 122, 4 (2017), 767–795. https://doi.org/10.1002/2016JG003640 arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2016JG003640

[11] Chongya Jiang and Youngryel Ryu. 2016. Multi-scale evaluation of global gross primary productivity and evapotranspiration products derived from Breathing Earth System Simulator (BESS). *Remote Sensing of Environment* 186 (2016), 528–547. https://doi.org/10.1016/j.rse.2016.08.030

[12] M. Jung, M. Reichstein, and A. Bondeau. 2009. Towards global empirical upscaling of FLUXNET eddy covariance observations: validation of a model tree ensemble approach using a biosphere model. *Biogeosciences* 6, 10 (2009), 2001–2013. https://doi.org/10.5194/bg-6-2001-2009

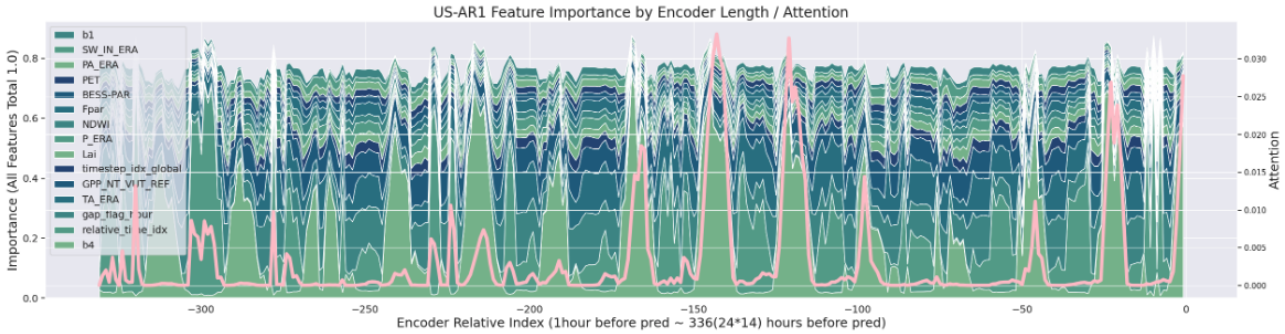


Figure 17: Feature importance with attention No-GPP-TFT model with 14 days encoder, full features Site: US-AR1, prediction time t: 12PM of August 8th, 2010

[13] Martin Jung, Markus Reichstein, Philippe Ciais, Sonia I. Seneviratne, Justin Sheffield, Michael L. Goulden, Gordon Bonan, Alessandro Cescatti, Jiquan Chen, Richard De Jeu, A. Johannes Dolman, Werner Eugster, Dieter Gerten, Damiano Gianelle, Nadine Gobron, Jens Heinke, John Kimball, Beverly E. Law, Leonardo Montagnani, Qiaozhen Mu, Brigitte Mueller, Keith Oleson, Dario Papale, Andrew D. Richardson, Olivier Roupsard, Steve Running, Enrico Tomelleri, Nicolas Viovy, Ulrich Weber, Christopher Williams, Eric F. Wood, Sönke Zaehle, and Ke Zhang. 2010. Recent decline in the global land evapotranspiration trend due to limited moisture supply. *Nature* 467, 7318 (21 oct 2010), 951–954. https://doi.org/10.1038/nature09396 Funding Information: Acknowledgements This work used eddy covariance data acquired by the FLUXNET community and in particular by the following networks: AmeriFlux (US Department of Energy, Biological and Environmental Research, Terrestrial Carbon Program; DE-FG02-04ER63917 and DE-FG02-04ER63911), AfriFlux, AsiaFlux, CarboAfrica, CarboEuropeIP, CarboItaly, CarboMont, ChinaFlux, Fluxnet-Canada (supported by the Canadian Foundation for Climate and Atmospheric Sciences, the Natural Sciences and Engineering Research Council of Canada, BIOCAP, Environment Canada and Natural Resources Canada), GreenGrass, KoFlux, the Largescale Biosphere–Atmosphere Experiment inAmazonia,theNordicCentrefor Studies ofEcosystemCarbonExchange, OzFlux, the Terrestrial Carbon Observatory System Siberia and US–China Carbon Consortium. We acknowledge the support to the eddy covariance data harmonization provided by CarboEuropeIP; the Food and Agriculture Organization of the United Nations' Global Terrestrial Observing System Terrestrial Carbon Observations; the Integrated Land Ecosystem–Atmosphere Processes Study, a core project of the International Geosphere–Biosphere Programme; the Max Planck Institute for Biogeochemistry; the National Science Foundation; the University of Tuscia; Université Laval; Environment Canada; and the US Department of Energy. We acknowledge database development and technical support from Berkeley Water Center, Lawrence Berkeley National Laboratory, Microsoft Research eScience, Oak Ridge National Laboratory, University of California, Berkeley and University of Virginia. We thank the members of FLUXNET (http://www.fluxdata.org/DataInfo) for their help with the data on this work. TRMM soil moisture retrievals and analysis were supported by the European Union (FP6) funded integrated project called WATCH (contract number 036946)thatsupportedA.J.D.and R.d.J.M.J.andM.R. were supportedbythe European Union (FP7) integrated project COMBINE (number 226520) and a grant from the Max-Planck Society establishing the MPRG Biogeochemical Model-Data Integration. C.W. was supported by the US National Science Foundation under grant ATM-0910766. D.P. acknowledges the support of the Euro–Mediterranean Centre for Climate Change. We acknowledge institutions and projects for free access to relevant data: the Global Runoff Data Centre, the Global Soil Wetness Project 2, the Global Precipitation Climatology Centre, the Global Precipitation Climatology Project, the Global Historical Climatology Network, the Potsdam Institute for Climate Impact Research, the University of East Anglia, the National Oceanic and Atmospheric Administration Earth System Research Laboratory and the European Centre for Medium-Range Weather Forecasts..

[14] Martin Jung, Markus Reichstein, Hank A. Margolis, Alessandro Cescatti, Andrew D. Richardson, M. Altaf Arain, Almut Arneth, Christian Bernhofer, Damien Bonal, Jiquan Chen, Damiano Gianelle, Nadine Gobron, Gerald Kiely, Werner Kutsch, Gitta Lasslop, Beverly E. Law, Anders Lindroth, Lutz Merbold, Leonardo Montagnani, Eddy J. Moors, Dario Papale, Matteo Sottocornola, Francesco Vaccari, and Christopher Williams. 2011. Global patterns of land-atmosphere fluxes of carbon dioxide, latent heat, and sensible heat derived from eddy covariance, satellite, and meteorological observations. *Journal of Geophysical Research: Biogeosciences* 116, G3 (2011). https://doi.org/10.1029/2010JG001566 arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2010JG001566

[15] Martin Jung, Markus Reichstein, Christopher R. Schwalm, Chris Huntingford, Stephen Sitch, Anders Ahlström, Almut Arneth, Gustau Camps-Valls, Philippe Ciais, Pierre Friedlingstein, Fabian Gans, Kazuhito Ichii, Atul K. Jain, Etsushi Kato, Dario Papale, Ben Poulter, Botond Raduly, Christian Rödenbeck, Gianluca Tramontana, Nicolas Viovy, Ying Ping Wang, Ulrich Weber, Sönke Zaehle, and Ning Zeng. 2017. Compensatory water effects link yearly global land CO 2 sink changes to temperature. *Nature* 541, 7638 (26 Jan. 2017), 516–520. https://doi.org/10.1038/nature20780 Funding Information: G.C.-V. was supported by the EU under ERC consolidator grant SEDAL-647423. Publisher Copyright: © 2017 Macmillan Publishers Limited, part of Springer Nature..

[16] C.O Justice, J.R.G Townshend, E.F Vermote, E Masuoka, R.E Wolfe, N Saleous, D.P Roy, and J.T Morisette. 2002. An overview of MODIS Land data processing and product status. *Remote Sensing of Environment* 83, 1 (2002), 3–15. https://doi.org/10.1016/S0034-4257(02)00084-6 The Moderate Resolution Imaging Spectroradiometer (MODIS): a new generation of Land Surface Monitoring.

[17] Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. 2020. Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting. arXiv:1912.09363 [stat.ML]

[18] Kaighin A McColl, Qing He, Hui Lu, and Dara Entekhabi. 2019. Short-term and long-term surface soil moisture memory time scales are spatially anticorrelated at global scales. *Geophysical Research Letters* 46, 8 (2019), 4352–4360.

[19] R.B Myneni, S Hoffman, Y Knyazikhin, J.L Privette, J Glassy, Y Tian, Y Wang, X Song, Y Zhang, G.R Smith, A Lotsch, M Friedl, J.T Morisette, P Votava, R.R Nemani, and S.W Running. 2002. Global products of vegetation leaf area and fraction absorbed PAR from year one of MODIS data. *Remote Sensing of Environment* 83, 1 (2002), 214–231. https://doi.org/10.1016/S0034-4257(02)00074-3 The Moderate Resolution Imaging Spectroradiometer (MODIS): a new generation of Land Surface Monitoring.

[20] J.E. Nash and J.V. Sutcliffe. 1970. River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology* 10, 3 (1970), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

[21] DARIO PAPALE and RICCARDO VALENTINI. 2003. A new assessment of European forests carbon exchanges by eddy fluxes and artificial neural network spatialization. *Global Change Biology* 9, 4 (2003), 525–535. https://doi.org/10.1046/j.1365-2486.2003.00609.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1046/j.1365-2486.2003.00609.x

[22] Markus Reichstein, Eva Falge, Dennis Baldocchi, Dario Papale, Marc Aubinet, Paul Berbigier, Christian Bernhofer, Nina Buchmann, Tagir Gilmanov, André Granier, Thomas Grünwald, Katka Havránková, Hannu Ilvesniemi, Dalibor Janous, Alexander Knohl, Tuomas Laurila, Annalea Lohila, Denis Loustau, Giorgio Matteucci, Tilden Meyers, Franco Miglietta, Jean-Marc Ourcival, Jukka Pumpanen, Serge Rambal, Eyal Rotenberg, Maria Sanz, John Tenhunen, Günther Seufert, Francesco Vaccari, Timo Vesala, Dan Yakir, and Riccardo Valentini. 2005. On the separation of net ecosystem exchange into assimilation and ecosystem respiration: review and improved algorithm. *Global Change Biology* 11, 9 (2005), 1424–1439. https://doi.org/10.1111/j.1365-2486.2005.001002.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2486.2005.001002.x

[23] Steven W. Running, Ramakrishna R. Nemani, Faith Ann Heinsch, Maosheng Zhao, Matt Reeves, and Hirofumi Hashimoto. 2004. A Continuous Satellite-Derived Measure of Global Terrestrial Primary Production. *BioScience* 54, 6 (06 2004), 547–560. https://doi.org/10.1641/0006-3568(2004)054[0547:ACSMOG]2.0.CO;2 arXiv:https://academic.oup.com/bioscience/article-pdf/54/6/547/26895742/54-6-547.pdf

[24] Youngryel Ryu, Dennis D. Baldocchi, Hideki Kobayashi, Catharine van Ingen, Jie Li, T. Andy Black, Jason Beringer, Eva van Gorsel, Alexander Knohl, Beverly E. Law, and Olivier Roupsard. 2011. Integration of MODIS land and atmosphere products with a coupled-process model to estimate gross primary productivity and evapotranspiration from 1 km to global scales. *Global Biogeochemical Cycles* 25, 4 (2011). https://doi.org/10.1029/2011GB004053 arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2011GB004053

[25] Youngryel Ryu, Chongya Jiang, Hideki Kobayashi, and Matteo Detto. 2018. MODIS-derived global land products of shortwave radiation and diffuse and total photosynthetically active radiation at 5km resolution from 2000. *Remote Sensing of Environment* 204 (2018), 812–825. https://doi.org/10.1016/j.rse.2017.09.021

[26] Crystal B Schaaf, Feng Gao, Alan H Strahler, Wolfgang Lucht, Xiaowen Li, Trevor Tsang, Nicholas C Strugnell, Xiaoyang Zhang, Yufang Jin, Jan-Peter Muller, Philip Lewis, Michael Barnsley, Paul Hobson, Mathias Disney, Gareth Roberts, Michael Dunderdale, Christopher Doll, Robert P d'Entremont, Baoxin Hu, Shunlin Liang, Jeffrey L Privette, and David Roy. 2002. First operational BRDF, albedo nadir reflectance products from MODIS. *Remote Sensing of Environment* 83, 1 (2002), 135–148. https://doi.org/10.1016/S0034-4257(02)00091-3 The Moderate Resolution Imaging Spectroradiometer (MODIS): a new generation of Land Surface Monitoring.

[27] G. Tramontana, M. Jung, C. R. Schwalm, K. Ichii, G. Camps-Valls, B. Ráduly, M. Reichstein, M. A. Arain, A. Cescatti, G. Kiely, L. Merbold, P. Serrano-Ortiz, S. Sickert, S. Wolf, and D. Papale. 2016. Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. *Biogeosciences* 13, 14 (2016), 4291–4313. https://doi.org/10.5194/bg-13-4291-2016

[28] Masahito Ueyama, Kazuhito Ichii, Hiroki Iwata, Eugénie S. Euskirchen, Donatella Zona, Adrian V. Rocha, Yoshinobu Harazono, Chie Iwama, Taro Nakai, and Walter C. Oechel. 2013. Upscaling terrestrial carbon dioxide fluxes in Alaska with satellite remote sensing and support vector regression. *Journal of Geophysical Research: Biogeosciences* 118, 3 (2013), 1266–1281. https://doi.org/10.1002/jgrg.20095 arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/jgrg.20095

[29] Zhengming Wan, Yulin Zhang, Qincheng Zhang, and Zhao liang Li. 2002. Validation of the land-surface temperature products retrieved from Terra Moderate Resolution Imaging Spectroradiometer data. *Remote Sensing of Environment* 83, 1 (2002), 163–180. https://doi.org/10.1016/S0034-4257(02)00093-7 The Moderate Resolution Imaging Spectroradiometer (MODIS): a new generation of Land Surface Monitoring.

[30] Wenhui Wang, Shunlin Liang, and Tilden Meyers. 2008. Validating MODIS land surface temperature products using long-term nighttime ground measurements. *Remote Sensing of Environment* 112, 3 (2008), 623–635. https://doi.org/10.1016/j.rse.2007.05.024

[31] Binrong Wu, Lin Wang, and Yu-Rong Zeng. 2022. Interpretable wind speed prediction with multivariate time series and temporal fusion transformers. *Energy* 252 (2022), 123990. https://doi.org/10.1016/j.energy.2022.123990

[32] Jingfeng Xiao, Jiquan Chen, Kenneth J. Davis, and Markus Reichstein. 2012. Advances in upscaling of eddy covariance measurements of carbon and water fluxes. *Journal of Geophysical Research: Biogeosciences* 117, G1 (2012). https://doi.org/10.1029/2011JG001889 arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2011JG001889

[33] Jingfeng Xiao, Qianlai Zhuang, Dennis D. Baldocchi, Beverly E. Law, Andrew D. Richardson, Jiquan Chen, Ram Oren, Gregory Starr, Asko Noormets, Siyan Ma, Shashi B. Verma, Sonia Wharton, Steven C. Wofsy, Paul V. Bolstad, Sean P. Burns, David R. Cook, Peter S. Curtis, Bert G. Drake, Matthias Falk, Marc L. Fischer, David R. Foster, Lianhong Gu, Julian L. Hadley, David Y. Hollinger, Gabriel G. Katul, Marcy Litvak, Timothy A. Martin, Roser Matamala, Steve McNulty, Tilden P. Meyers, Russell K. Monson, J. William Munger, Walter C. Oechel, Kyaw Tha Paw U, Hans Peter Schmid, Russell L. Scott, Ge Sun, Andrew E. Suyker, and Margaret S. Torn. 2008. Estimation of net ecosystem carbon exchange for the conterminous United States by combining MODIS and AmeriFlux data. *Agricultural and Forest Meteorology* 148, 11 (2008), 1827–1847. https://doi.org/10.1016/j.agrformet.2008.06.015

[34] Jingfeng Xiao, Qianlai Zhuang, Beverly E. Law, Jiquan Chen, Dennis D. Baldocchi, David R. Cook, Ram Oren, Andrew D. Richardson, Sonia Wharton, Siyan Ma, Timothy A. Martin, Shashi B. Verma, Andrew E. Suyker, Russell L. Scott, Russell K. Monson, Marcy Litvak, David Y. Hollinger, Ge Sun, Kenneth J. Davis, Paul V. Bolstad, Sean P. Burns, Peter S. Curtis, Bert G. Drake, Matthias Falk, Marc L. Fischer, David R. Foster, Lianhong Gu, Julian L. Hadley, Gabriel G. Katul, Roser Matamala, Steve McNulty, Tilden P. Meyers, J. William Munger, Asko Noormets, Walter C. Oechel, Kyaw Tha Paw U, Hans Peter Schmid, Gregory Starr, Margaret S. Torn, and Steven C. Wofsy. 2010. A continuous measure of gross primary production for the conterminous United States derived from MODIS and AmeriFlux data. *Remote Sensing of Environment* 114, 3 (2010), 576–591. https://doi.org/10.1016/j.rse.2009.10.013

[35] Xiangming Xiao, Stephen Boles, Jiyuan Liu, Dafang Zhuang, and Mingliang Liu. 2002. Characterization of forest types in Northeastern China, using multitemporal SPOT-4 VEGETATION sensor data. *Remote Sensing of Environment* 82, 2 (2002), 335–348. https://doi.org/10.1016/S0034-4257(02)00051-2

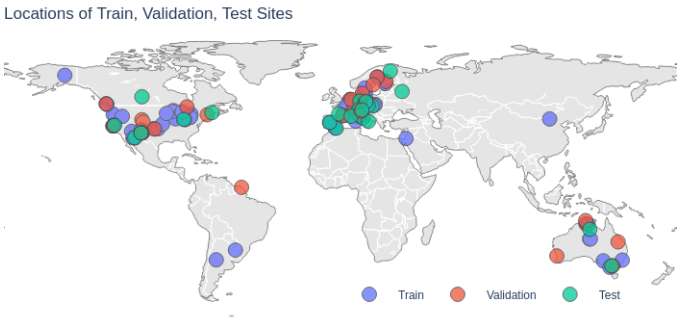


Figure 18: Locations of the train, validation, test datasets used in this study.

[36] Feihua Yang, Kazuhito Ichii, Michael A. White, Hirofumi Hashimoto, Andrew R. Michaelis, Petr Votava, A-Xing Zhu, Alfredo Huete, Steven W. Running, and Ramakrishna R. Nemani. 2007. Developing a continental-scale measure of gross primary production by combining MODIS and AmeriFlux data through Support Vector Machine approach. *Remote Sensing of Environment* 110, 1 (2007), 109–122. https://doi.org/10.1016/j.rse.2007.02.016

[37] Y. Zhang, J. Joiner, S. H. Alemohammad, S. Zhou, and P. Gentine. 2018. A global spatially contiguous solar-induced fluorescence (CSIF) dataset using neural networks. *Biogeosciences* 15, 19 (2018), 5779–5800. https://doi.org/10.5194/bg-15-5779-2018

[38] Maosheng Zhao, Faith Ann Heinsch, Ramakrishna R. Nemani, and Steven W. Running. 2005. Improvements of the MODIS terrestrial gross and net primary production global data set. *Remote Sensing of Environment* 95, 2 (2005), 164–176. https://doi.org/10.1016/j.rse.2004.12.011

A LOCATIONS OF TRAIN/VALIDATION/TEST SITES

Since this study modeled without taking on an dependency on the geolocation of each flux tower site, the data split is not dependent on location either. Interestingly, the stratified split still resulted a fairly diverse set of datasets. In Figure 18, the validation dataset in red has the more sites outside of Europe and North American than the train and test dataset. The timelines distribution of the data split is shown in Figure **??**.

B FEATURE ENGINEERING OF TREE MODELS

C CORRELATIONS OF ORIGINAL FEATURES

D VALIDATION METRICS

The below table displays the model performance metrics on validation split.

No.	Model	Features/Params	Train	Encoder	Decoder	RMSE*	MAE*	NSE*
–	**Baseline**							
1	RFR-BASELINE	Original /Untuned	5 year	-	-	3.840	2.220	0.723
2	XGB-BASELINE	Original /Untuned	5 year	-	-	3.607	2.008	0.743
–	**Feature Engineering RFR/XGB**							
3	RFR-ENG	Original + Engineered	5 year	-	-	3.878	2.292	0.718
4	XGB-ENG	Original + Engineered	5 year	-	-	TBD	TBD	TBD
5	RFR-TOP9	Top 9 features*/Tuned	5 year	-	-	3.584	1.992	0.759
6	XGB-TOP3	Top 3 features*/Tuned	5 year	-	-	3.665	2.048	0.748
–	**Benchmark TFT with Past GPP**							
7	GPP-TFT-7E1T-U	Original/Untuned	1 year	24*7	1	2.503	1.245	0.850
8	GPP-TFT-1E1T	Original/Tuned	1 year	24*1	1	2.336	1.163	0.898
9	GPP-TFT-3E1T	Original/Tuned	1 year	24*3	1	2.298	1.142	0.901
10	GPP-TFT-7E1T	Original/Tuned	1 year	24*7	1	2.281	1.124	0.903
11	GPP-TFT-14E1T	Original/Tuned	1 year	24*14	1	2.98	1.167	0.899
–	**Upscaling Tree-FT**							
12	RFR-TFT-14D	Slim Features*/Tuned	5 year	24*14	1	3.563	1.945	0.759
13	XGB-TFT-14D	Slim Features*/Tuned	5 year	24*14	1	3.670	1.985	0.744
–	**Upscaling No-GPP-TFT**							
14	No-GPP-TFT-3D-16HS	Slim Features*/Untuned	5 year	24*3	1	3.660	2.034	0.745
15	No-GPP-TFT-7D-16HS	Slim Features*/Untuned	5 year	24*7	1	3.636	2.030	0.748
16	No-GPP-TFT-7D-64HS	Slim Features*/Untuned	5 year	24*7	1	3.658	2.029	0.746
17	No-GPP-TFT-14D-16HS	Slim Features*/Untuned	5 year	24*14	1	3.618	2.001	0.751
18	No-GPP-TFT-30D-16HS	Slim Features*/Untuned	5 year	24*30	1	3.792	2.022	0.739

Table 6: Comparison of model Result of RFR, XGBoost and TFT models
***TOP9: SW-IN-ERA, hour, VPD-ERA, NIRv, NDVI, EVI, TA-ERA. TOP3: NDVI, NIRv, SW-IN-ERA. Slim Features: TA-ERA, SW-IN-ERA, LW-IN-ERA, VPD-ERA, P-ERA, PA-ERA, NDVI, b2, b4, b6, b7, BESS-PARdiff, CSIF-SIFdaily, ESACCI-smi, Percent-Snow, LAI, LST-Day, LST-Night**

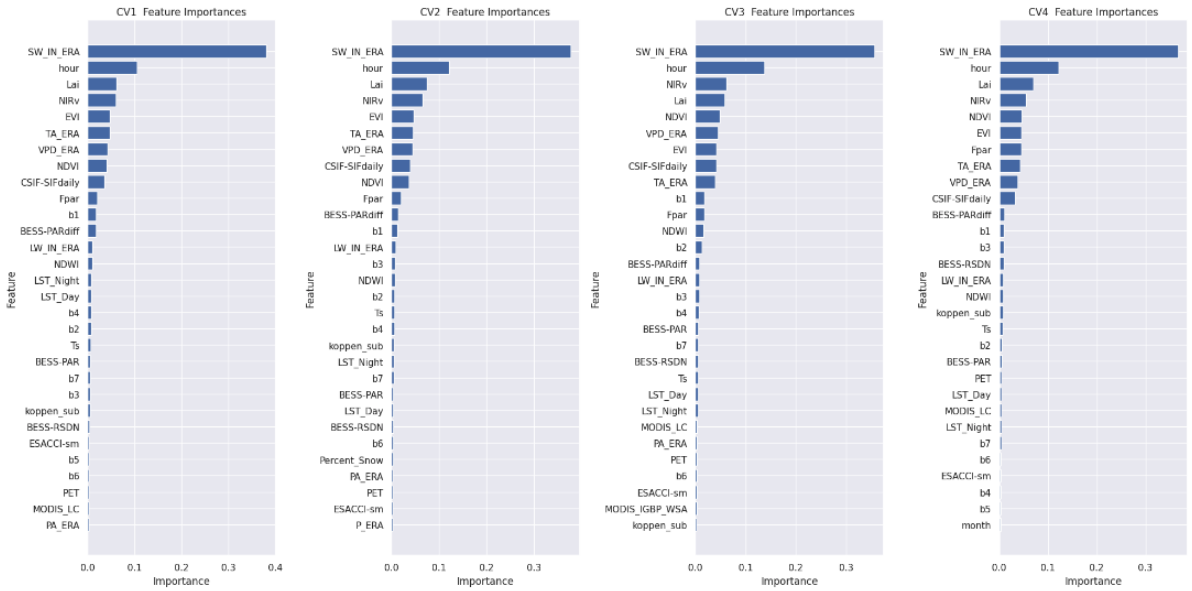


Figure 20: Feature importances of RFR

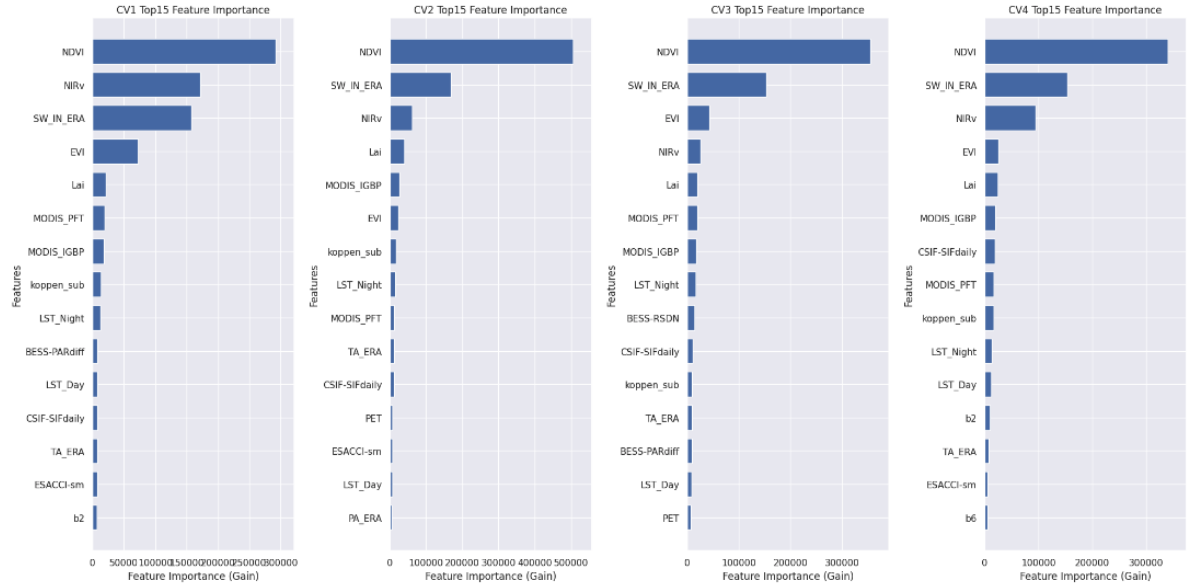
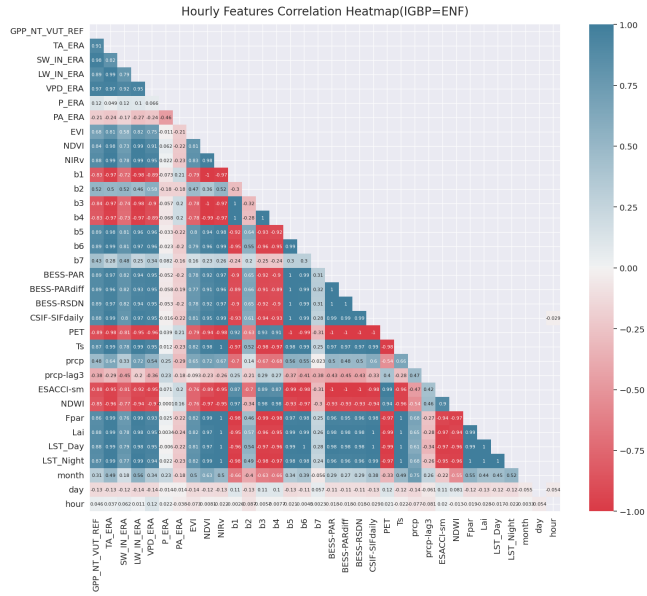
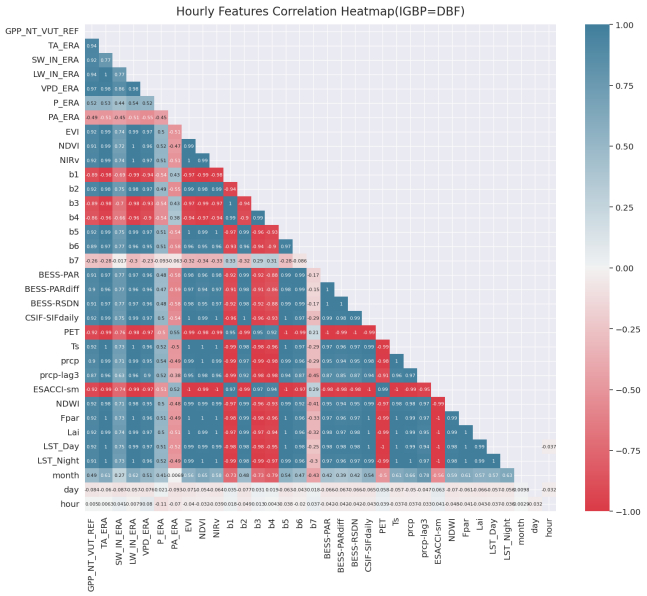


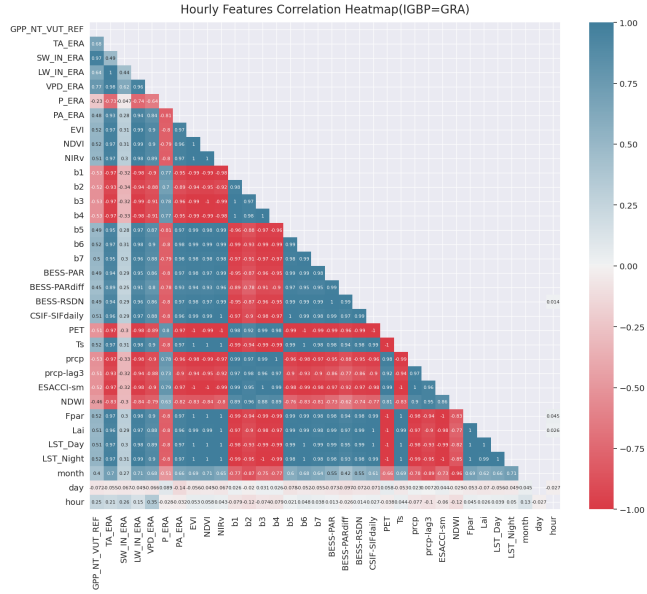
Figure 21: Feature importances of XGBoost



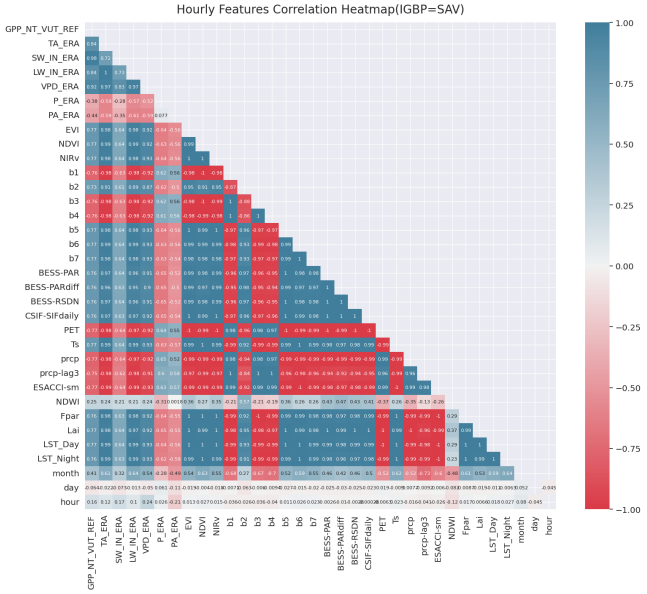
(a) IGBP ENF



(b) IGBP DBF



(c) IGBP GRA



(d) IGBP SAV

Figure 22: Examples of Correlations by IGBP Type