Upscaling Global Hourly GPP with Temporal Fusion Transformer (TFT)

Rumi Nakagawa* Mary Chau* John Calzaretta* Trevor Keenan Puya Vahabi
Alberto Todeschini Maoya Bassiouni Yanghui Kang

University of California, Berkeley
{ruminakagawa, marychau, john.calzaretta, trevorkeenan, puyavahabi,
todeschini, maoya, yanghuikang}@berkeley.edu

Abstract

Reliable estimates of Gross Primary Productivity (GPP), crucial for evaluating climate change initiatives, are currently only available from sparsely distributed eddy covariance tower sites. This limitation hampers access to reliable GPP quantification at regional to global scales. Prior machine learning studies on upscaling in situ GPP to global wall-to-wall maps at sub-daily time steps faced limitations such as lack of input features at higher temporal resolutions and significant missing values. This research explored a novel upscaling solution using Temporal Fusion Transformer (TFT) without relying on past GPP time series. Model development was supplemented by Random Forest Regressor (RFR) and XGBoost, followed by the hybrid model of TFT and tree algorithms. The best preforming model yielded to model performance of 0.704 NSE and 3.54 RMSE. Another contribution of the study was the breakdown analysis of encoder feature importance based on time and flux tower sites. Such analysis enhanced the interpretability of the multi-head attention layer as well as the visual understanding of temporal dynamics of influential features.

1. Introduction

The ongoing increase in CO_2 emissions, despite a temporary slowdown during the 2020 pandemic, has mobilized government agencies, industries, and individuals to tackle climate change. As investments in climate change mitigation initiatives continue to grow, reliable estimation of carbon flux becomes essential for informed green policies, global carbon budgets, and ensuring the effectiveness and accountability of these actions. Gross Primary Productivity (GPP), which refers to the quantification of carbon uptake through vegetation, is a key factor in the carbon cycle and estimated from CO_2 exchanges measurements. However, the measurement of CO_2 exchanges is currently limited to a sparse network of flux towers, tall instrumented structures used to quantify gas exchange between the land surface and atmosphere of the Earth. With less than a thousand flux towers globally, the availability of carbon flux measurements is geographically limited. Given the extreme sparsity of sites with GPP measurements, the development of an upscaling model is imperative to estimate GPP at locations worldwide. This upscaling task is particularly challenging due to regional biases in the available flux data, mainly concentrated in North America and Europe. Improving empirical models for global inference will enhance the accessibility of quantified CO_2, regardless of the availability of flux tower equipment.

Past literature that studied global GPP products has shifted its focus to higher temporal resolutions, from annual [1], monthly [8] [9], daily [20], and to the latest study on half-hourly [3] level. Estimating GPP at sub-daily or hourly time scales is crucial for understanding important ecosystem-climate interactions, such as the lagged or legacy effects, especially during extreme climate events, which are projected to become more frequent under climate change [22] [32]. However, the model with half-hourly resolution encountered the limited availability of features of the same granularity. In the latest study, meteorological features were constrained to daily resolution as the highest level of detail [3]. Moreover, machine learning studies in this field have not adequately captured the temporal dynamics, partially due to the computational resources required for time-aware models, such as long short-term memory (LSTM).

Temporal Fusion Transformer (TFT) possesses key characteristics that address these limitations: 1) TFT accommodates time series of different time periods. 2) TFT can forecast on previously unseen entities. 3) TFT handles diverse inputs, including heterogeneous time series, time-varying features, and static metadata. 4) TFT utilizes LSTM and self-attention mechanism to learn historical patterns in both

*These authors contributed equally.

short- and long-term. 5) TFT offers powerful interpretability. Built-in function of Pytorch allows feature importance analysis by the encoder, decoder and static features at each prediction time step. Overall, TFT potentially serves as an effective upscaling solution, and provides rich information to comprehend the behavior of the model.

This study applied TFT to upscale hourly GPP by incorporating time-aware elements. The primary objectives were to enhance the performance of the upscaling model and to analyze the temporal dynamics of influential features in the TFT model output. Additionally, Random Forest Regressor (RFR) and XGBoost (XGB) models were developed in parallel using the same dataset. Notably, three modeling approaches were employed. First, a non-upscaling capable model with past GPP values was constructed to establish the best possible performance benchmark for the TFT model. Second, a TFT model was built without the past GPP values of the test sites to serve as an upscaling solution. Finally, a two-stage model, TFT models was developed utilizing the predicted GPP values of the tree algorithms as potential alternatives of the unavailable past GPP values. It aimed to assess whether estimated past GPP values could improve to model performance than not providing any past GPP values.

2. Literature Review

The FLUXNET project, initiated in 1997 to develop accessible ground truth, was continued with the cooperation of scientific communities to develop networks across Europe, North America, and Asia. The first empirical model that used machine learning applied artificial neural networks to data in European forests [17]. Support Vector Regression [30] [21] [7] and ensemble tree models [28] [29] [8] [3] also have been actively studied. Tramontana et al. compared 16 machine learning algorithms, including kernel methods, neural networks, tree methods, and regression splines [20] and reported high consistency in the model performances, although site-dependencies were observed. While recurrent neural networks, such as LSTM was applied to CO_2 flux prediction as a time series forecasting model [2], its application to GPP has not been explored in previous literature.

Satellite-based remote sensing features are commonly used in this field due to their geographically widespread availability and ability to indicate vegetation structural changes [28] [29] [27]. Past studies utilized remote sensing features such as Land Surface Temperature (LST) [24] [25], Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI) [6], Leaf Area Index (LAI), fraction of absorbed photosynthetic active radiation (fPAR) [14], Normalized Difference Water Index (NDWI) [4]. In addition to remote sensing features, meteorological features such as air temperature (TA), vapor pressure deficit (VPD), water availability [10] were also often included in the past studies [20] [3].

GPP upscaling with higher temporal resolution has gained attention in the past two decades. Previous studies by Jung et al. in 2009 and 2011 focused on monthly GPP [8] [9], while Beer et al. examined median annual GPP in 2010 [1], and Tramontana et al. explored daily GPP in 2016 [20], followed by Bodesheim et al. utilizing Random Forest Regressor (RFR) for half-hourly predictions in 2018 [3]. Bodesheim et al. applied the remote sensing and meteorological features previously utilized by Tramontana et al. Although the model targets were at a half-hourly level, most features, except one radiation feature, were on daily level, resulting in common values for every 48 time steps. To overcome this limitation, Bodesheim et al. employed two approaches: 1) Building a model for each of the 48 time steps, 2) Constructing a single model with the first order derivative of target value (GPP) as a supplemental engineered feature. The latter approach resulted better model performance, with an NSE value of 0.67.

3. Data Sources

In order to build a global GPP upscaling model and fully utilize the ability of TFT in handling heterogeneous inputs, this study incorporated features from multiple global data sources, encompassing remote sensing and meteorological data at different temporal resolutions, including hourly, daily, 4-day, 8-day, 16-day, and monthly intervals. Additionally, climate and land classification labels for each location were included as static metadata. The target variable, GPP, was directly obtained from the FLUXNET2015 Dataset [18].

3.1. GPP Ground-Truth: FLUXNET

FLUXNET is a valuable source of original and processed data collected from flux towers, with the latest data product, FLUXNET2015 Dataset, hosted by Lawrence Berkeley National Laboratory. Carbon fluxes are measured through flux towers using the eddy covariance method, which quantifies gas exchange between the biosphere and the atmosphere. FLUXNET offers four types of GPP based on the timing of respiration measurements, and the calculation methods for thresholds. GPP-NT-VUT-REF was defined as the target GPP variable in this study. The duration of available GPP data varies drastically across sites, ranging from over 20 years to just a few weeks (Figure 1).

3.2. Global Features: Remote-Sensing Data

MODIS (Moderate Resolution Imaging Spectroradiometer) is the key instrument carried by the Terra (EOS AM-1) and Aqua (EOS PM-1) satellites. These satellites capture images of the surface of the Earth every day or every two days, providing data with a spatial resolution of 500 m to 1 km [11]. The study used the following MODIS datasets with different resolutions and features:

1. MCD43C4 (Daily): NIRv (Near-Infrared Reflectance of Vegetation[1]), NDVI, EVI, NDWI, percentage of snow, PET (Evapotranspiration), Surface reflectance b1 to b7 [19]
2. MCD15A3H (4-day): LAI (Leaf Area Index), fPAR (Fraction of Photosynthetically Active Radiation) [15]
3. MCD12Q1 (Annual): MODIS-IGBP (International Geosphere–Biosphere Programme) and PFT (Plant Functional Type) [5]
4. MYD11A1 (Daily): Daytime LST (Land Surface Temperature), Nighttime LST [23]

Two other global remote sensing data were applied in this study: 1) Continuous Solar-Induced Fluorescence (CSIF) with 4-day resolution derived from Orbiting Carbon Observatory-2 (OCO-2) SIF observations and MODIS surface reflectance. Inclusion of CSIF potentially allows models to capture more details on the diurnal and seasonal patterns of photosynthesize activities. 2) Photosynthetically active radiation (PAR), diffuse PAR, and shortwave downwelling radiation (RSDN) derived from Breathing Earth System Simulator (BESS) models. These BESS_Rad data helps account for variations in light availability and quality that can affect photosynthesis in different vegetation types and canopy structures, thereby potentially contributing to estimates of GPP.

3.3. Global Features: Meteorological Data

Two notable meteorological and climate data sources in the study are ERA5-Land and Köppen-Geiger classification. ERA5-Land is a dataset from the European Centre for Medium-Range Weather Forecasts (ECMWF), offering hourly climate and global environmental data on land surface [13]. Air temperature, VPD, precipitation, skin temperature, soil moisture, potential evapotranspiration, shortwave and longwave radiation from this dataset are used as features in this study. The Köppen classification represents the major climate groups base on annual temperature and precipitation patterns while Köppen-sub is a subcategory with more details on seasonality and precipitation characteristics.

4. Models and Method

This study focused on applying a temporal fusion transformer (TFT), supplemented by RFR and XGBoost regression to develop GPP upscaling models.

Temporal Fusion Transformer (TFT), introduced by Google Cloud AI in 2019 [12], is an attention-based time series forecasting model that has outperformed existing models in various time series forecasting tasks [12] [26]. The unique components in TFT architecture make it a powerful and interpretable model for time series forecasting. It applies a static covariate encoder to integrate static metadata into the temporal fusion decoder, enabling temporal forecasts conditioned on such information. Its gating components allow the model to skip unnecessary parts of the network, improving model efficiency. Variable selection is applied to both time-variant and time-invariant input features, selecting relevant features and removing noisy inputs to enhance model performance. Moreover, the model learns short- and long-term patterns through sequence-to-sequence (LSTM) and attention-based layers, identifying temporal dynamics such as seasonality and significant events.

Given I unique entities in a time series data, each entity i has associated 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 are subdivided into two categories: 1) observed inputs $\boldsymbol{z}_{i,t} \in \mathbb{R}^{m_x}$ that are known prior time t but unknown afterward 2) predetermined inputs $\boldsymbol{x}_{i,t} \in \mathbb{R}^{m_x}$ which remains known after time t (e.g. the hour of the day) . 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 incorporates the past target values and observed values within a finite look-back window of length k (i.e., $y_{i,t-k:t} = y_{i,t-k}, ..., y_{i,t}$), followed by known input features across the entire time range (i.e., $\boldsymbol{x}_{i,t-k:t+\tau} = \boldsymbol{x}_{i,t-k}, ..., \boldsymbol{x}_{i,t}, ..., \boldsymbol{x}_{i,t+\tau}$), to produce forecasts for τ_{max} time steps ahead.

In the context of reconstructing past global GPP, each location (e.g. flux tower) corresponded to an entity i. All input features except GPP measurements were available in prediction time t as well as τ steps ahead. Remote sensing and meteorological data, varying by time, were considered as time-varying known features. Likewise, time-related information such as month, day, and hour of the record fell in the same input type as well. The metadata of each location, such as IGBP and Köppen labels, were treated as static covariates. Since the study focused on single-point estimate of the past GPP at unseen sites, the decoder length (τ) of the TFT models were consistently set to 1. The encoder length (k) was a hyperparameter to be optimized to balance model complexity and performance.

4.1. Evaluation metrics

Three evaluation metrics were selected to assess the performance of the GPP upscaling models and to compare with the past literature: 1) Root Mean Squared Error (RMSE), 2) Mean Absolute Error (MAE), and 3) Nash-Sutcliffe Efficiency (NSE). NSE is a normalized statistic that determines the relative magnitude of the residual variance in comparison with the measured data variance [16]. Although origi-

[1] canopy structure that are measured in remote sensing

Figure 1. Timelines of available observations of all flux towers (274 total) with GPP measurement, organized by IGBP groups. Each horizontal line corresponds to a single flux tower site, and colored area depicts the available period of data in each site.

nally developed for hydrology, the NSE is commonly employed in upscaling tasks and is mathematically equivalent to the coefficient of determination (R^2) [9] [20] [3].

5. Data Preprocessing

To accommodate time and computational constraints, the modeling period for the data was set from 2010 to 2015, during which a greater number of observations and relevant sites were accessible. As depicted in Figure 1, the number of active flux towers significantly decreased after 2015. In order to mitigate the impact of gap-filling and to preserve the year-long seasonality in the dataset, the flux towers with less than a year of available data and/or over 20% of missing records were excluded. The resulting dataset applied to the modeling comprised 129 flux tower sites.

5.1. Imputation and Gap-Filling

After the initial filtering of the flux tower sites, the dataset still had a considerable amount of missing values and gaps (i.e. time steps with no records). The gaps in record sequences present a challenge for time series models, such as TFT, that require continuous sequences of input data.

K-nearest neighbors (KNN) imputation was employed to address the challenge of large amount of missing data. This approach considers temporal dependencies, fills large blocks of missing values, and maintains time series continuity by identifying similar neighbors based on Euclidean distance, thereby making it well-suited to address the problem at hand. Two KNN imputation methods were used in the study: one for filling missing values within time step records and another for filling gaps in record sequences. Both methods utilized 5 neighbors, Euclidean distance, and a uniform average of neighbor values, determined through testing and validation with the RFR model. Gap-filling flags were included as observed input features in all TFT experiments to indicate whether a record was gap-filled or not.

5.2. Stratified Train/Test Split

Since the number of flux towers is limited, it is vital to maintain diversity in training, validation, and test sets to ensure generalization and objective evaluation of the model. In order to realize the diverse distribution of each dataset, stratified data split was employed by generic IGBP groups. It combined some categories, such as Open and Closed Shrublands into Shrublands, and Woody Savannas and Savannas into Savannas. This stratified approach resulted in 78 sites for training, 26 sites for validation, and 25 sites for testing.

6. Experimental design

Step	Experiment	Algorithms	Upscaling Capable
1	Baseline	RFR	Yes
2	Feature-Engineered Trees	RFR XGBoost	Yes
3	GPP-TFT	TFT	No
4	No-GPP-TFT	TFT	Yes
5	Tree-FT	Hybrid (Tree + TFT)	Yes

Table 1. Summary of 5-step experimental design.

The experimental design consisted of the 5 steps shown in Table 1. The first step involved building a baseline was

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 2. RFR-BASELINE model performance of each CV group.

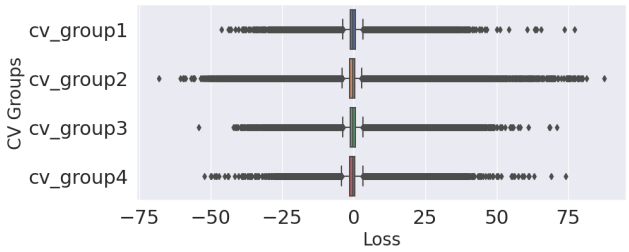

Figure 2. Distribution of loss of each CV group on RFR-BASELINE.

built using RFR and full features. Cross-validation (CV) was conducted with RFR to assess the potential volatility in model performance on different combinations of sites in training, validation, and testing sets. Due to the substantial resource demands associated with TFT model development accompanied by CV, the impact of relying a specific split group for model validation was investigated using the baseline RFR model. Subseuqently, feature engineering was applied to the tree models (RFR, XGBoost) to study whether certain features contributed significantly to the performance of non-time-aware models, followed by the dimensionality reduction. The output of tuned tree models was utilized in the hybrid model for subsequent experiments.

The initial experiment on the GPP-TFT model followed the default TFT usage, assuming the ideal scenario where past GPP values were available at prediction time. Though the ultimate goal of the study was upscaling, a scenario where historical GPP measurements would not be available at locations without flux towers, the best-possible benchmark model was developed to understand the upper limit of the TFT model performance. The Tree-FT models substituted historical GPP input with estimated past GPP values from the tree models, while the No-GPP-TFT models were developed without any past GPP provided as input features. Additionally, the study also explored temporal dynamics of model behavior by analyzing interpretable outputs of the TFT model.

7. Results

Data prepocessing with stratified train/test split divided the flux tower sites into five groups. Reserving one group for testing, 4-fold CV were performed on the tree models. RFR model performance (Table 2) and distributions of loss (Figure 2) across the CV groups showed negligible differences among the four groups, suggesting that using one specific CV group as validation set throughout the study may be sufficient for objective model evaluation. Based on these findings, the TFT models was developed with consistently assigning fourth CV group as the validation set.

The results of the best models after hyperparameter tuning are presented in Table 3. Model 1 was the baseline using RFR, and Model 2 and 3 implemented dimensionality reduction, with each algorithm having the optimal number of features. Model 4, GPP-TFT, included past GPP values as a predictor feature. Models 5 and 6 were Tree-FT models, which were upscaling models that utilized the predicted GPP values from RFR and XGBoost as estimated past GPP values. Model 7, the No-GPP-TFT model, represented the upscaling approach without any past GPP value.

In summary, the best-performing upscaling model was the RFR model with the top 9 features (Model 2) with 3.423 RMSE and 0.704 NSE. The GPP-TFT model, though not capable of upscaling, achieved 2.132 RMSE And 0.886 NSE, implying the potential performance ceiling for global hourly GPP upscaling models. As a result of hyperparameter tuning and feature engineering, Tree-FT and No-GPP-TFT models, both capable of upscaling, also demonstrated improvements when compared to the past literature.

8. TFT Model Interpretability

TFT offers interpretability beyond the feature importance provided in tree models. It allows for explorations of feature significance over time and provides attention information that identifies the most influential encoder time step at the site level. By leveraging the interpretable model output of the TFT model, model behavior were analyzed by IGBP groups, and a novel visual-analytical approach was applied to investigate the temporal dynamics of influential features in each prediction.

8.1. Analysis by IGBP Groups

Achieving a highly generalized and accurate GPP inference model across diverse global regions presents challenges. Evaluation metric breakdown by IGBP groups on the No-GPP-TFT model (Table 4) shows variations in model performance. Deciduous Broadleaf Forests (DBF) and Grasslands (GRA) show better performance than the original model(Model 7 in Table 3), while some IGBP groups, such as Evergreen Broadleaf Forests (EBF) and Shrublands (OSH), exhibit underperformance. The attention plot for selected IGBP groups (GRA, OSH, and WET) in Figure 3 indicates potentially distinct patterns among IGBP groups. OSH displays a relatively flat attention curve

Upscaling Model	Features	Hidden Size	Encoder Length	Decoder Length	RMSE	MAE	NSE
Past Literature							
0. Bodesheim 2018(RFR)	-	-	-	-	3.940	-	0.670
Baseline							
1. RFR-BASELINE	Original	-	-	-	3.675	2.011	0.678
Feature-Engineered RFR/XGB							
2. RFR-TOP9	Top 9 features[1]	-	-	-	3.523	1.841	0.704
3. XGB-TOP3	Top 3 features[2]	-	-	-	3.610	1.870	0.677
GPP-TFT(non-upscaling)							
4. GPP-TFT-14EN-ORG[3]	Original	136	24*14	1	2.132	1.016	0.886
Tree-FT							
5. RFR-TFT-14EN-SLIM	Slim Features[4]	16	24*14	1	3.630	1.900	0.671
6. XGB-TFT-14EN-SLIM	Slim Features	16	24*14	1	3.807	2.002	0.638
No-GPP-TFT							
7. No-GPP-TFT-7EN-SLIM	Slim Features	16	24*7	1	3.594	1.904	0.677

Table 3. Upscaling model results of the test set.
[1]TOP9: SW-IN-ERA (Shortwave Radiation), NDVI, NIRv, hour, LAI, TA-ERA, VPD-ERA (Vapor Pressure Deficit), EVI, CSIF-SIFdaily
[2]TOP3: NDVI, NIRv, SW-IN-ERA
[3]Limited to one year of training data due to resource constraints.
[4]Slim Features: TA-ERA, SW-IN-ERA, LW-IN-ERA (Longwave Radiation), VPD-ERA, P-ERA, PA-ERA, NDVI, b2, b4, b6, b7, BESS-PARdiff, CSIF-SIFdaily, ESACCI-smi, Percent-Snow, LAI, LST-Day, LST-Night

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

Table 4. No-GPP-TFT model evaluation metrics breakdown by IGBP groups (ordered by NSE values).

that peaks around 24 hours prior, while GRA and WET show distinct daily variations peaking around 18 hours ahead. These findings suggest potential distinctive trends and differential processing of historical information based on land cover types during inference. Additionally, Figure 3 highlights that attention moves in different directions over time; WET has upward trend, whereas GRA and OSH show downward trends of attention. These observations implies the importance of considering varying timescales for various regions and selecting an appropriate encoder length for the better performance of the inference.

8.2. Temporal Analysis of Feature Importance

There were two reasons for identifying encoder variable importance as the scope of the analysis: 1) The raw output of encoder/decoder variable importance allowed for further breakdown and examination. 2) Since the study defined decoder length as one, the resulted decoder attention was zero for all the experiments. Going beyond the default PyTorch API usage, model dynamics and behavior were analyzed by

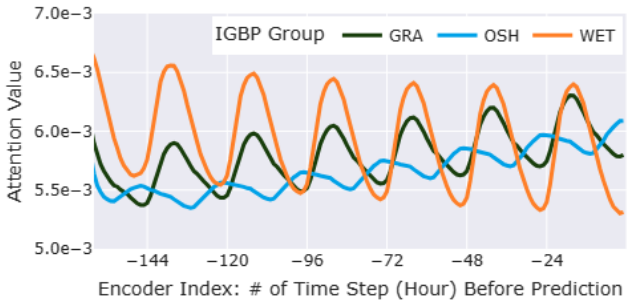

Figure 3. Attention of selected IGBP groups: Grasslands (GRA), Open Shrublands (OSH), and Wetlands (WET), obtained from the No-GPP-TFT model.

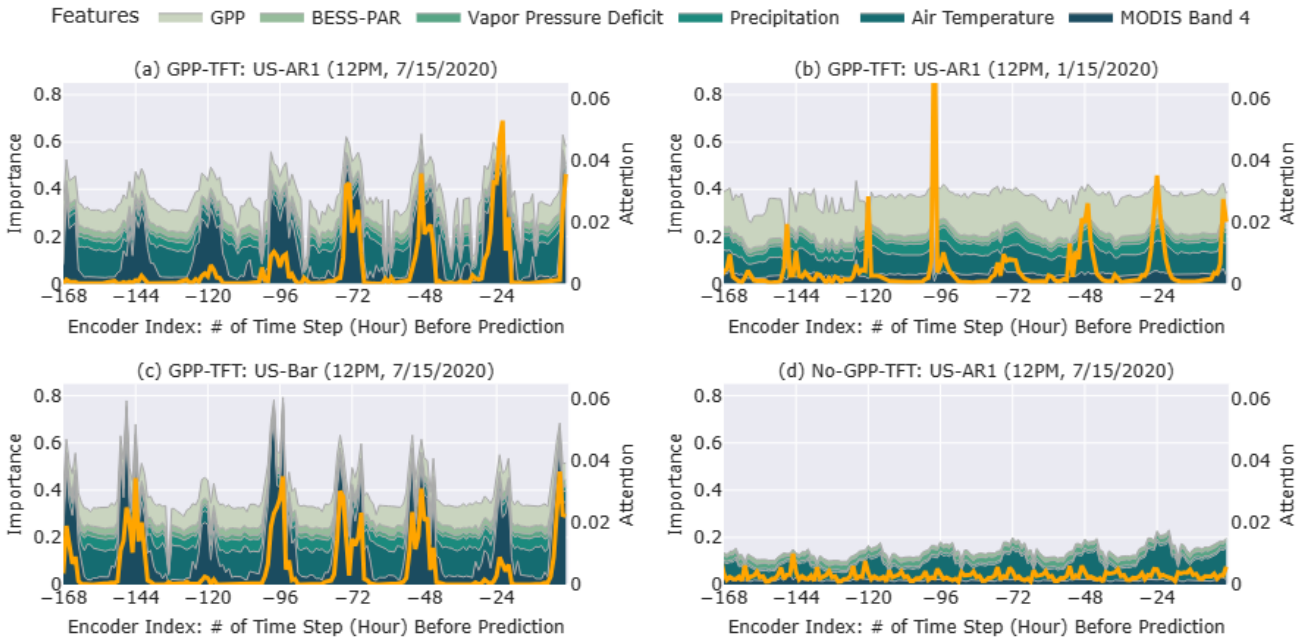

Figure 4. Temporal analysis of encoder feature importance value by encoder index on selected features. (a) GPP-TFT: Prediction at US-AR1 on July 15, 2020, at 12 PM (b) GPP-TFT: Prediction at US-AR1 on January 15, 2020, at 12 PM (c) GPP-TFT: Prediction at US-Bar on July 15, 2020, at 12 PM (d) No-GPP-TFT: Prediction at July 15, 2020, at 12 PM

visualizing temporal allocation of feature importance. The visual provided a detailed breakdown of snapshots at each prediction time and site, which had not been extensively investigated before. By recognizing the pivotal role of encoder variables in shaping encoder attention patterns, the study meticulously plotted the evolving feature importance across the relative time indices of the encoder length.

8.2.1 Visual Analysis Walk-Through

Figure 4a provides a snapshot of the encoder feature importance at site US-AR1[2] of the GPP-TFT model (Model 4 in Table 3) on July 15, 2020, with a prediction time of 12 PM. The x-axis represents hours leading up to the prediction time t, with negative values indicating hours before t (Ex. -24 represents 24 hours prior to t). The left y-axis shows the feature importance value, while the right y-axis represents the attention at each encoder index. The attention of the snapshot is depicted as a bold yellow line, and the feature importance of selected features is displayed in a stacked plot format. While the Figure 4 only covers six selected sample features, the importance are added up to one as a total of all the features in each encoder index. In more details, the top 15 influential features are listed in Table 5.

In this snapshot, MODIS spectral band 4 (b4; darkest green) stood out as the most influential feature. It exhibits spikes every 24 hours before the prediction time t (noon in July in the northern hemisphere), while features such as air temperature (TA-ERA) peaked during the evening. The largest spikes in overall attention coincided with the spikes in b4 implies its influence in predicting GPP at t. The feature importance of b4 even surpasses the importance of past GPP measurements during the daytime. These patterns of b4 that GPP-TFT model captured potentially aligns properties of vegetation in biology; The relation of b4 with photosynthetic activities of vegetation [31].

8.2.2 Comparative Analysis of Seasonality, IGBP Groups and Models

Comparative analysis of temporal dynamic of encoder feature importance were examined from three perspectives: 1) Seasonality 2) IGBP Groups 3) GPP-TFT versus No-GPP-TFT (Model 4 and 7 in Table 3). When comparing the visualization of encoder feature importance between summer and winter (Figure 4a and Figure 4b), significant differences are observed in the shape and order of feature. In winter, b4 shows less impact than summer, while past GPP and air temperature exhibit increased influence.

The feature importance was also compared at two locations representing different IGBP groups: US-AR1 (GRA, Grasslands) and US-Bar[3] (DBF, Deciduous Broadleaf

[2] ARM USDA UNL OSU Woodward Switchgrass 1, Oklahoma, USA

[3] Bartlett Experimental Forest, New Hampshire, USA

TFT Model	GPP-TFT			No-GPP
Site	US-AR1		US-Bar	US-AR1
Month	July	Jan	July	July
Subplots of Figure 4	(a)	(b)	(c)	(d)
1. b4	11.7	3.9	14.1	2.3
2. Hourly Gap Flag	11.4	9.5	12.3	9.4
3. Relative Time Index	10.4	12.6	5.4	1.0
4. GPP	8.5	15.0	7.9	-
5. Air Temperature	8.2	8.3	8.5	7.3
6. Global Time Index	3.2	1.0	2.7	2.9
7. Precipitation	3.1	4.7	3.6	1.0
8. LAI	3.0	1.8	2.1	2.1
9. NDWI	2.8	2.9	2.8	2.3
10. fPAR	2.6	3.6	2.0	2.9
11. BESS-PAR	2.4	2.0	2.5	1.5
12. Shortwave Radiation	2.3	1.3	2.5	7.7
13. PA	2.3	1.9	2.9	2.3
14. PET	2.2	0.9	1.9	0.5
15. b1	2.0	3.0	1.9	1.8

Table 5. Top 15 average encoder feature importance (%) of snapshots in Figure 4. Features are ordered by the rank of Figure 4a, GPP-TFT model prediction at US-AR1 on July 15, 2020, at 12 PM)

Forests) (Figure 4a and Figure 4c). The feature importance of b4 of the two locations show a shared diurnal pattern on the same prediction date and time. Similar patterns are also observed at FI-Hyy[4] (ENF, Evergreen Needleleaf Forests) on the same prediction timestamp, and AU-DaP[5] (GRA, Grasslands) on January 15, 2020 (i.e. summer at Australia). Despite the distinct climate and vegetation among different IGBP groups, similar temporal dynamics is observed. On the other hand, the model exhibits distinct trend in attention by IGBP (Yellow line in Figure 4a and Figure 4c). In US-AR1, the attention reaches the peak 24 hours before the prediction time, follow by the diminishing daily trend. It implies that day before the prediction time is the most contributing information of the prediction. On the flip side, attention is relatively evenly distributed in US-Bar, while disregarding information in one and five days prior. These differences indicate that the TFT model has the potential to capture both shared and location-specific temporal dynamics in feature importance for each prediction.

Furthermore, a comparison was made between the temporal encoder feature importance of the GPP-TFT model (Figure 4a) and the No-GPP-TFT model (Figure 4d). The key distinction between the two models lies in whether the model incorporates past GPP values as input features or not, resulting in significant differences in both model performance (Table 3) and the ranking of influential features (Table 5). The selected six features in the No-GPP-TFT model (Figure 4d) are limited to approximately 20% of the total importance in the snapshot. Shortwave radiation accounts for 7.7% in the No-GPP-TFT model, whereas b4, the most impactful feature in summer snapshot of GPP-TFT model(Figure 4a (Figure 4c), is limited to 2.3% of the importance. Similarly, relative time index, the potential clue for capturing temporal patterns, represents 1.0% of importance in the No-GPP-TFT model, while it surpasses 10% in the GPP-TFT model. These differences show the varying decision-making processes and model priorities based on the model setup.

[4]Hyytiala, Finland

[5]Daly River Savanna, Australia

9. Conclusion

This study addressed the crucial need for reliable estimation of GPP worldwide. Time-aware models leveraging the Temporal Fusion Transformer (TFT) model were developed to estimate hourly GPP values. Several modeling approaches were explored, including establishing a benchmark performance of TFT models with past GPP values as predictors, building TFT models without past GPP as input for upscaling, and developing two-stage models that integrated predictions from the tree models into TFT models. Comparing to the benchmark performance established in the previous work, improved performance was observed in several models. The best-performing model, feature-engineered RFR, achieved NSE of 0.704 and RMSE of 3.54.

The extensive analysis of model performance by IGBP groups showed widely distributed performance, implying distinct model dynamics by vegetation types. Instead of relying on a single common model for all geographical locations, training distinct models for each unique IGBP group might enhance the limited performance of the TFT model.

The successful implementation of TFT in an unconventional application, solving time series forecasting problems without past target values, lays the groundwork for future upscaling solutions using TFT. Furthermore, the novel visual-analytical approach derived from interpretable outputs from TFT provides valuable insights into the temporal dynamics of feature importance, extending its potential beyond carbon flux research. These findings contribute to the advancement of GPP estimation, and hold promise for potential applications in other industries.

References

[1] 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. Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate. *Science*, 329(5993):834–838, 2010. 1, 2

[2] 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. Memory effects of climate and vegetation affecting net ecosystem co2 fluxes in global forests. *PLOS ONE*, 14(2):1–22, 02 2019. 2

[3] P. Bodesheim, M. Jung, F. Gans, M. D. Mahecha, and M. Reichstein. Upscaled diurnal cycles of land–atmosphere fluxes: a new global half-hourly data product. *Earth System Science Data*, 10(3):1327–1365, 2018. 1, 2, 4

[4] Bo cai Gao. Ndwi—a normalized difference water index for remote sensing of vegetation liquid water from space. *Remote Sensing of Environment*, 58(3):257–266, 1996. 2

[5] Mark Friedl and Damien Sulla-Menashe. Mcd12q1 modis/terra+aqua land cover type yearly l3 global 500m sin grid v006, 2019. 3

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

[7] 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. 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):767–795, 2017. 2

[8] M. Jung, M. Reichstein, and A. Bondeau. Towards global empirical upscaling of fluxnet eddy covariance observations: validation of a model tree ensemble approach using a biosphere model. *Biogeosciences*, 6(10):2001–2013, 2009. 1, 2

[9] 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. 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. 1, 2, 4

[10] 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. Compensatory water effects link yearly global land co 2 sink changes to temperature. *Nature*, 541(7638):516–520, Jan. 2017. 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. 2

[11] C.O Justice, J.R.G Townshend, E.F Vermote, E Masuoka, R.E Wolfe, N Saleous, D.P Roy, and J.T Morisette. An overview of modis land data processing and product status. *Remote Sensing of Environment*, 83(1):3–15, 2002. The Moderate Resolution Imaging Spectroradiometer (MODIS): a new generation of Land Surface Monitoring. 2

[12] Bryan Lim, Sercan O. Arik, Nicolas Loeff, and Tomas Pfister. Temporal fusion transformers for interpretable multi-horizon time series forecasting, 2020. 3

[13] J Muñoz Sabater. Era5-land monthly averaged data from 1950 to present, Jul 2023. 3

[14] 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. Global products of vegetation leaf area and fraction absorbed par from year one of modis data. *Remote Sensing of Environment*, 83(1):214–231, 2002. The Moderate Resolution Imaging Spectroradiometer (MODIS): a new generation of Land Surface Monitoring. 2

[15] Ranga Myneni, Yuri Knyazikhin, and Taejin Park. Mcd15a3h modis/terra+aqua leaf area index/fpar 4-day l4 global 500m sin grid v006, 2015. 3

[16] J.E. Nash and J.V. Sutcliffe. River flow forecasting through conceptual models part i — a discussion of principles. *Journal of Hydrology*, 10(3):282–290, 1970. 3

[17] DARIO PAPALE and RICCARDO VALENTINI. A new assessment of european forests carbon exchanges by eddy fluxes and artificial neural network spatialization. *Global Change Biology*, 9(4):525–535, 2003. 2

[18] Gilberto Pastorello, Carlo Trotta, Eleonora Canfora, Housen Chu, Danielle Christianson, You-Wei Cheah, Cristina Poindexter, Jiquan Chen, Abdelrahman Elbashandy, Marty Humphrey, et al. The fluxnet2015 dataset and the oneflux processing pipeline for eddy covariance data. *Scientific data*, 7(1):1–27, 2020. 2

[19] Crystal Schaaf and Zhuosen Wang. Mcd43c4 modis/terra+aqua brdf/albedo nadir brdf-adjusted ref daily l3 global 0.05deg cmg v006, 2015. 3

[20] 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. Predicting carbon dioxide and energy fluxes across global fluxnet sites with regression algorithms. *Biogeosciences*, 13(14):4291–4313, 2016. 1, 2, 4

[21] Masahito Ueyama, Kazuhito Ichii, Hiroki Iwata, Eugénie S. Euskirchen, Donatella Zona, Adrian V. Rocha, Yoshinobu Harazono, Chie Iwama, Taro Nakai, and Walter C. Oechel. Upscaling terrestrial carbon dioxide fluxes in alaska with satellite remote sensing and support vector regression. *Journal of Geophysical Research: Biogeosciences*, 118(3):1266–1281, 2013. 2

[22] J. von Buttlar, J. Zscheischler, A. Rammig, S. Sippel, M. Reichstein, A. Knohl, M. Jung, O. Menzer, M. A. Arain, N. Buchmann, A. Cescatti, D. Gianelle, G. Kiely, B. E. Law, V. Magliulo, H. Margolis, H. McCaughey, L. Merbold, M. Migliavacca, L. Montagnani, W. Oechel, M. Pavelka, M. Peichl, S. Rambal, A. Raschi, R. L. Scott, F. P. Vaccari, E. van Gorsel, A. Varlagin, G. Wohlfahrt, and M. D. Mahecha. Impacts of droughts and extreme-temperature events on gross primary production and ecosystem respiration: a systematic assessment across ecosystems and climate zones. *Biogeosciences*, 15(5):1293–1318, 2018. 1

[23] Zhengming Wan, Simon Hook, and Glynn Hulley. Myd11a1 modis/aqua land surface temperature/emissivity daily l3 global 1km sin grid v006, 2015. 3

[24] Zhengming Wan, Yulin Zhang, Qincheng Zhang, and Zhao liang Li. Validation of the land-surface temperature products retrieved from terra moderate resolution imaging spectroradiometer data. *Remote Sensing of Environment*, 83(1):163–180, 2002. The Moderate Resolution Imaging Spectroradiometer (MODIS): a new generation of Land Surface Monitoring. 2

[25] Wenhui Wang, Shunlin Liang, and Tilden Meyers. Validating modis land surface temperature products using long-term nighttime ground measurements. *Remote Sensing of Environment*, 112(3):623–635, 2008. 2

[26] Binrong Wu, Lin Wang, and Yu-Rong Zeng. Interpretable wind speed prediction with multivariate time series and temporal fusion transformers. *Energy*, 252:123990, 2022. 3

[27] Jingfeng Xiao, Jiquan Chen, Kenneth J. Davis, and Markus Reichstein. Advances in upscaling of eddy covariance measurements of carbon and water fluxes. *Journal of Geophysical Research: Biogeosciences*, 117(G1), 2012. 2

[28] 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. Estimation of net ecosystem carbon exchange for the conterminous united states by combining modis and ameriflux data. *Agricultural and Forest Meteorology*, 148(11):1827–1847, 2008. 2

[29] 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. A continuous measure of gross primary production for the conterminous united states derived from modis and ameriflux data. *Remote Sensing of Environment*, 114(3):576–591, 2010. 2

[30] 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. 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):109–122, 2007. 2

[31] Gaofei Yin, Aleixandre Verger, Adrià Descals, Iolanda Filella, and Josep Peñuelas. A broadband green-red vegetation index for monitoring gross primary production phenology. *Journal of Remote Sensing*, 2022, 2022. 7

[32] M. Zhang and X. Yuan. Rapid reduction in ecosystem productivity caused by flash droughts based on decade-long fluxnet observations. *Hydrology and Earth System Sciences*, 24(11):5579–5593, 2020. 1