Temporal and Spatial Net Ecosystem Exchange (NEE) Prediction

Sweta Bhattacharya
University of California, Berkeley
Information and Data Science

Joshua Dunn
University of California, Berkeley
Information and Data Science

Marcia(Yiying) Liu
University of California, Berkeley
Information and Data Science

1 Abstract

This research aims to predict Net Ecosystem Exchange (NEE) across temporal and spatial dimensions using remote sensing, climate, and eddy-covariance (EC) flux datasets. The study focuses on 185 forest and woodland FLUXNET sites globally, analyzing the response of disturbance and climate variations on Gross Primary Productivity (GPP) and Net Ecosystem Exchange (NEE) over time. The research involves data preprocessing, feature engineering, model selection, and training to predict NEE accurately. The results are visualized through graphs and charts, providing insights into the patterns and trends of NEE over time and space. The study is compelling and impactful as it contributes to the understanding of carbon sequestration through the fast cycle of photosynthesis and helps reduce total CO2 in the ecosystem. Additionally, the data produced by the models can be used in new and ongoing research to understand and mitigate the effects of global climate change.

2 Introduction

The global climate is changing rapidly, posing severe risks to the sustainability of our planet's ecosystem. One of the primary concerns is the accumulation of greenhouse gases, such as carbon dioxide, in the atmosphere, leading to an increase in the Earth's temperature. Forests and woodlands play a vital role in mitigating this threat by absorbing and storing carbon through photosynthesis. However, climate variations and disturbances such as droughts, fires, and deforestation impact the productivity and carbon uptake of these ecosystems. Therefore, understanding the response of these ecosystems to environmental changes is crucial in mitigating the effects of climate change. This study aims to predict Net Ecosystem Exchange (NEE) across temporal and spatial dimensions using remote sensing, climate, and eddy-covariance (EC) flux datasets.

The study focuses on 185 forest and woodland FLUXNET sites globally, analyzing the response of disturbance and climate variations on Gross Primary Productivity (GPP) and NEE over time. The results of this study can provide valuable insights into the patterns and trends of NEE, contributing to the understanding of carbon sequestration through the fast cycle of photosynthesis and helping to reduce the total CO2 in the ecosystem. Additionally, the data produced by the models can be used in new and ongoing research to understand and mitigate the effects of global climate change.

3 Model Baseline Comparison

3.1 XGBoost

We used XGBoost for our baseline model. Initially, we used this model to train the dataset in time. All data before 2018 were used for training and 2019 dataset for validation. We used TimeSeriesSplit for 5 fold cross validation and incorporated the hyperparameter tuning with GridSearchCV. The temporal plots for different IGBP's from the XGBoost model shows an R-square of 0.76 for "ENF" and R-square of 0.64 for shrublands (OSH and CSH).

3.1.1

Number of training data from shrublands, Savannah and wetlands are typically less than the other IGBPs.The reason for the low R-square for those IGBPs such as mixed forest (MF), croplands (CRO) can be investigated in future. The scientific community was more interested in finding out how the baseline XGBoost model will predict for a new location $site\_id$, for which we implemented a spatial XGBoost model. In the spatial XGBoost model, we first divided the dataset by IGBP for US sites. Despite the reduction in the number of training data, this approach was more meaningful to the scientific community because a comparative study can be done

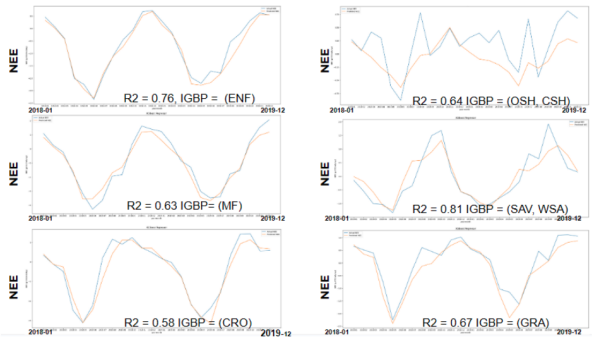

Fig. 1. NEE Prediction of Selected IGBPs

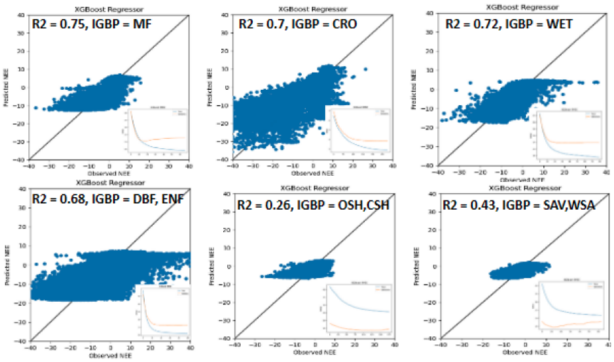

Fig. 2. NEE Prediction Distribution.

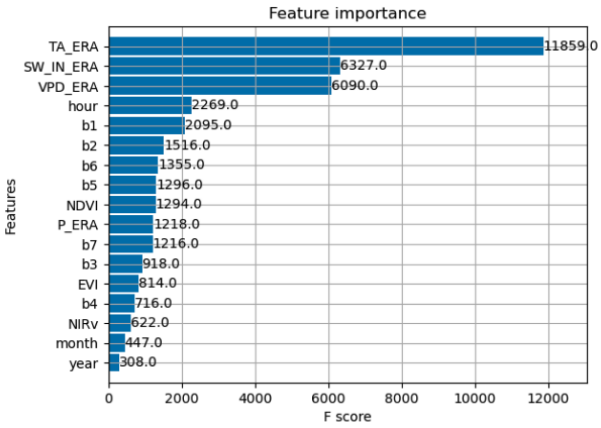

Fig. 3. Feature Importance

between the state of the art results versus our baseline model results. The sites with IGBP as MF, CRO, GRA, ENF, DBF and WET have more training data and generally developed a better model than the other IGBP's. However, in some cases there were so few sites (for example, for Savannah there were only 1 training site and one validation site),that the model shows underfitting.

3.1.2

The feature importance plot shows that the target variable NEE is more dependent on the climate variables such as Temperature, solar radiation and vapor pressure than the spectral bands or other time related features.

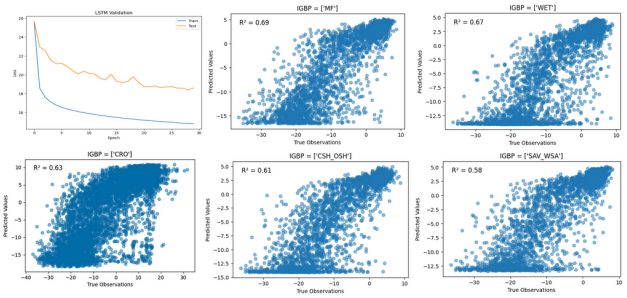

Fig. 4. Scatterplot of True Observations vs. Predicted Values for IGBP (Selected IGBPs)

3.2 LSTM

In this LSTM session, we aimed to model and predict the NEE_VUT_REF variable for different IGBP types using the Long Short-Term Memory (LSTM) approach. To achieve this, we first preprocessed the data for each IGBP type, combining some of them into combined categories (e.g., 'CSH_OSH' and 'SAV_WSA'). We then performed data cleaning and feature selection for each IGBP type, removing unwanted columns and handling missing values.

3.2.1 Data Preprocessing and Model Training

After preprocessing, we split the data into training and testing sets, and normalized the target variable. We then prepared the data for time-series modeling by creating a SequenceDataset class, which returns sequences of input data and their corresponding target values. DataLoader instances were created for both training and testing datasets.

We defined an LSTM_FORCAST model with one LSTM layer and a linear output layer. We trained the model using the Adam optimizer and Mean Squared Error (MSE) loss function. During training, we recorded train and test losses for each epoch to analyze the model's performance.

3.2.2 Model Evaluation and Results

After training the model, we used it to make predictions on the test dataset and evaluated the performance using metrics such as mean squared error, mean absolute error, and R-squared score. We plotted the predicted values against the true observations in a scatter plot to visualize the model's performance for each IGBP type. Additionally, we plotted the train and test losses during the training process to assess the model's learning progress and identify potential overfitting or underfitting.

In summary, this LSTM session demonstrated the application of an LSTM model to predict NEE_VUT_REF values for various IGBP types. We preprocessed the data, trained the model, and evaluated its performance using different metrics, presenting the results in graphical form for each IGBP type.

3.3 TFT

We use Pytorch Forecasting's implementation of Temporal Fusion Transformer for modeling along with the Py-

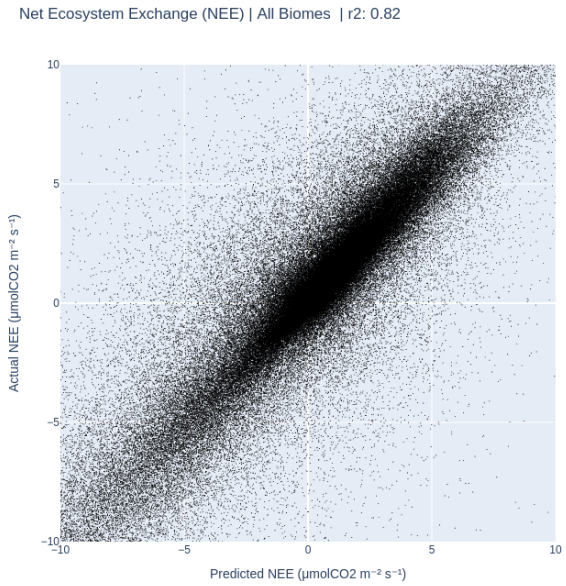

Fig. 5. Scatterplot of True Observations vs. Predicted Values for TFT (All IGBPs)

torch Lighting TimeSeriesDataset for managing data. The data was split by SITE_ID by random choice into a 70/30 train/test ratio. Site sampling was restricted to a single continous 1.5 year block to provide balancing due to the uneven distribution of samples per site. TFT allows for designation of known, unknown, static, and time-varying variables. The prediction horizon was limited to a single time step and so all variables were known. The model considered up to three days of historical half-hour measurements for training and prediction. Any missing values in the data were filled using a forward-fill strategy. We trained the model on an Amazon Web Services EC2 g4dn.16xlarge instance to utilize GPU-accelerated processing.

After many iterations and hyperparameter tuning, we settled on a TFT model with 30% dropout, 4 LSTM layers, 64 hidden layer nodes, 4 attention layers, and 32 hidden continuous nodes. The best-performing model's weights, along with all supporting codes, are saved to the accompanying code repository.

The TFT model performed exceptionally well with an R-square value of 0.82 across all IGBPs in the testing data.

3.4 Further Work

The TFT model has an interpretable multi-head attention matrix and can be analyzed to provide insight on the model's decision behavior and feature focus. A limitation of the FLUXNET dataset is geographic locality due to the limited measurement range of the eddy covariance towers. Further work should be done with the TFT model to make NEE predictions in areas where no sensors are present.

4 Conclusions and Discussions

In this study, we focused on the prediction of Net Ecosystem Exchange (NEE) in both temporal and spatial domains, aiming to improve the understanding of carbon dynamics across different ecosystems and to investigate the performance of various machine learning models in predicting NEE. We utilized a comprehensive dataset containing multiple IGBP land cover types and their corresponding NEE measurements.

We employed three different machine learning models, namely XGBoost, LSTM, and TFT, for the temporal prediction of NEE. Our analysis showed that the TFT architecture performed the best, outperforming both XGBoost and LSTM models. Therefore, we suggest the TFT model as the new state-of-the-art for NEE prediction. The superior performance of the TFT model can be attributed to its ability to account for both temporal and non-linear relationships between variables, capturing complex patterns in the data. The multi-head attention layer of the TFT model is interpretable, which allows for further exploration of the model's feature selection in future work.

The FluxNet data used in this study is geographically sparse in many continents, and where present, it is restricted to a very small radius of observation. The TFT model can potentially be extended through more work to predict NEE where flux tower data is not available. Incorporating more biophysical variables, such as soil moisture and land use, could further improve the prediction accuracy of NEE. Moreover, the use of remote sensing data, such as satellite-derived vegetation indices, can provide additional information on the spatial distribution of ecosystems and improve the model's generalizability.

In conclusion, this study highlights the importance of temporal and spatial modeling for predicting NEE and provides insights into the performance of different machine learning models in this context. The findings of this research can be valuable for understanding the carbon dynamics of ecosystems and informing decision-makers about the potential of machine learning in environmental monitoring and management. By suggesting the TFT model as the new state-of-the-art for NEE prediction and proposing potential extensions and improvements, this study contributes to the ongoing efforts to enhance our understanding and predictive capabilities of ecosystem processes.

5 Links

The GitHub repository, along with the trained model, can be found at: https://github.com/jdunns-organization/capstone

Table 1. Scores for Spatial Prediction (train/test models split based on IGBPs)

IGBP	**Number of Sites**	**Number of training samples**	**XGB - R^2**	**LSTM - R^2**	**TFT - R^2**
MF (mixed Forest)	4	116,353	0.75	0.61	0.57
GRA (Grassland)	9	576,534	0.43	0.52	0.87
ENF (Evergreen Needle Leaf)	17	1,199,243	0.66	0.43	0.77
CRO (Croplands)	5	279,361	0.7	0.63	0.89
CSH, OSH (shrublands)	3	379,442	0.26	0.61	0.73
SAV, WSA (Savannah)	7	84,529	0.43	0.58	0.78
WET (Wetlands)	4	253,728	0.72	0.67	0.79
All Biomes	-	-	0.70	0.60	0.82