Improved terrestrial Gross Primary Productivity (GPP) estimation using Multisource Data

---

A Report submitted to the
Indian Statistical Institute, Kolkata
for the award of the degree

of

Master of Statistics

by

Hemant Banke (MD2107)

Under the guidance of

Dr. B. Uma Shankar, MIU

Statistics and Mathematics Unit
Indian Statistical Institute Kolkata
West Bengal - 700 108, India

Table of Contents

Abstract

Understanding the carbon uptake of vegetation from the atmosphere is one of the most important indicators for predicting the future trend of climate change and for implementing sustainable development strategies. Terrestrial Gross Primary Productivity (GPP), an indicator of carbon uptake, denotes the total amount of carbon assimilated in terrestrial ecosystem through photosynthesis.

Flux tower based eddy covariance measurements are considered reliable estimates for GPP. But setting up a tower is a costly affair and hard to maintain, so we have limited number of towers for flux measurements. To upscale GPP from tower to regional and global scale, we can utilize satellite based measurements because of its higher resolution and continuous coverage and build model for estimation.

Machine Learning (ML) tools are very useful for building GPP models. ML tools have proven utility in complex nonlinear problem solving in the field of ecology. Therefore, we apply a data driven approach to predict GPP at any given location. The models will be built using flux tower based eddy covariance dataset and Remote Sensing data in combination with meteorological and topographical data for the Australian Region. The aim of this project is to predict GPP using machine learning models and understand it's dependencies on biophysical, meteorological and topographical features.

1 Introduction

Gross Primary Productivity (GPP) is a fundamental ecological concept that measures the amount of carbon fixed by plants through photosynthesis in a given area or ecosystem over a period of time, without taking into account any losses due to respiration or other factors. Essentially, GPP represents the total amount of energy that plants capture from the sun and convert into organic matter through photosynthesis. This process forms the foundation of most terrestrial ecosystems, and plays a crucial role in the global carbon cycle. GPP is an important metric for understanding ecosystem health and productivity, and can provide insights into the impacts of environmental changes such as climate change and land use on ecosystem functioning. It is widely studied and measured by ecologists, climatologists, and other scientists interested in understanding the functioning and dynamics of natural systems.

GPP is an important measure of the productivity of an ecosystem because it reflects the amount of energy that is available to support the growth and survival of other organisms within that system. In general, ecosystems with higher GPP are able to support a greater diversity of organisms and a higher biomass (i.e., the total mass of living organisms within an ecosystem) than those with lower GPP.

Another important forest carbon flux indicator usually concerned in the literature is Net Ecosystem Exchange (NEE). The CO_2 fluxes into or out of plants are referred to as NEE. Because most forests act as carbon sinks, absorbing CO_2 from the atmosphere, NEE for forests is frequently negative. However, in some parts of the planet, trees may act as carbon sinks, releasing more carbon than they take in, resulting in NEE being positive. Modelers of climate change are consequently very interested in NEE as a representation of net carbon dynamics.

The measurement of Gross Primary Productivity (GPP) is a complex and multi-faceted process, involving a variety of techniques and methods. The eddy covariance approach is used to

continuously quantify carbon flux exchange between ecosystem and atmosphere in terms of Net Ecosystem Exchange (NEE) in a conventional Flux tower. The NEE is then subdivided into GPP and ecosystem respiration (RE), with the proper respiration assumption taken into consideration. There are more than 600 flux tower stations worldwide. The eddy covariance method is generally considered to be the most accurate way to quantify GPP, despite the associated uncertainties. This is because it directly measures the exchange of CO_2 and water vapor between the ecosystem and the atmosphere, providing a high-resolution picture of the carbon balance of the ecosystem. However, the eddy covariance method requires specialized instrumentation and expertise, and is limited to relatively small spatial scales.

A variety of data-driven strategies can be used to upscale eddy covariance-based data from isolated flux towers. These include purely statistical techniques (like regression models and semi-empirical regression models), machine learning techniques (like neural networks and decision trees), and strategies based on the inherent water-use efficiency model [1] and the LUE model [2]–[4]. All of these strategies' model performance is subpar and is reliant on the environment as well as the availability of sufficient and accurate data [5]. Moreover, these models are not capable of temporal extrapolation (i.e., future predictions). Hence, we explore GPP estimation utilising ML on Remote Sensing data in combination with meteorological and topographical data as also done in [6].

2 Data Description

2.1 Sources for GPP

Two significant global datasets of terrestrial carbon flux are present in the literature.

One is the FLUXNET dataset [7] (https://fluxnet.org/data/), which uses eddy covariance techniques to quantify the fluxes of carbon, water, and energy between the biosphere and atmosphere. Estimates of the GPP are also provided. The dataset is recognised as the benchmark for modelling and validation that allows for the most precise in-situ measurement. But it suffers due to the limited distribution of the sites.

The MODIS GPP [8] (Moderate Resolution Imaging Spectroradiometer (MODIS)) and NPP product, or MOD17, is the other dataset (https://modis.gsfc.nasa.gov/data/). MODIS is mounted on the NASA Earth Observing System (EOS), which comprises of two polar-orbiting spacecraft, Terra (morning flyover) and Aqua (early afternoon flyover). Unfortunately, MOD17 statistics do not exactly correspond with FLUXNET data due to the algorithm's complexity, the large number of input variables, and a number of other factors.

2.2 Data for Australian Region

This study is based on the Australian Region as plenty of data is available in this region for training a good model. It's 7.692 million km^2 landmass is located between latitudes 9^oS and 44^oS and longitudes 112^oE and 154^oE. It exhibits a variety of landscapes, including tropical rainforests, mountain ranges, and deserts.

The data used for modelling was obtained as follows :

1. Biophysical Features (Obtained from MODIS Remote Sensing dataset):

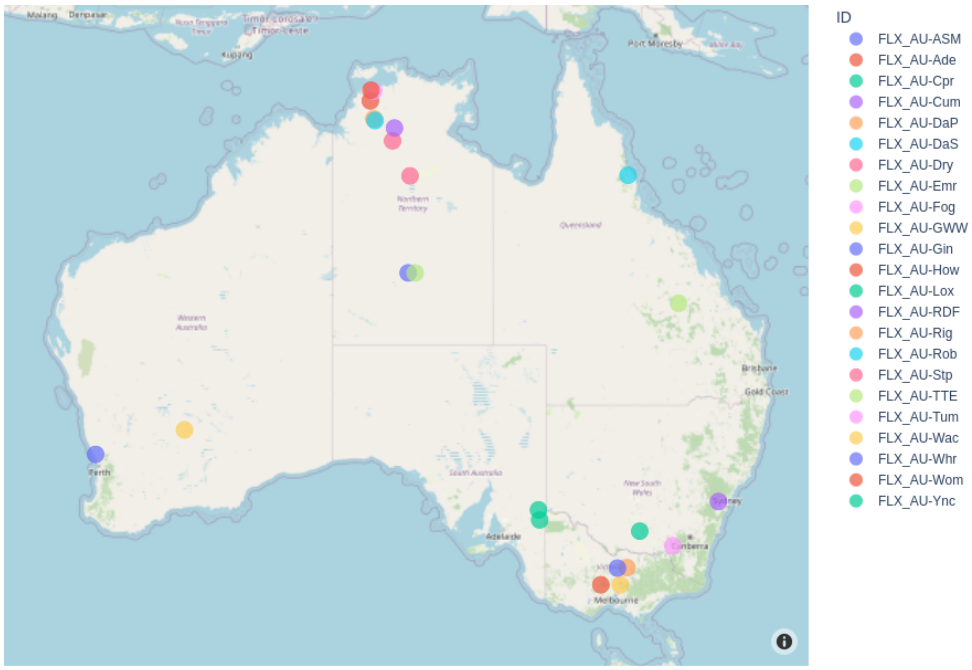


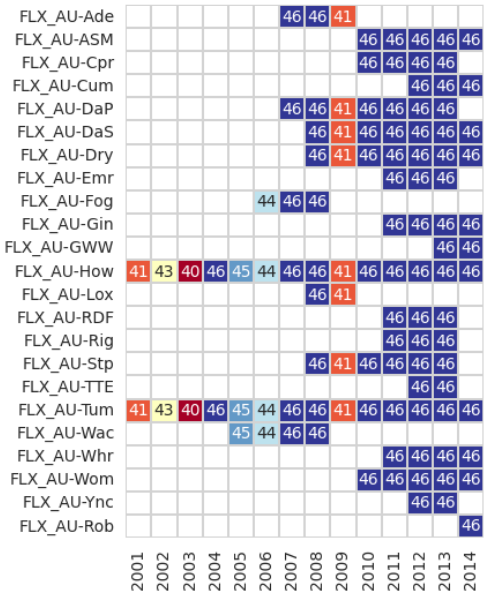
Figure 1: Map of the 23 Flux tower sites in Australia

Enhanced Vegetation Index (EVI) : quantifies vegetation greenness, also corrects for some atmospheric conditions and canopy background noise
Leaf Area Index (LAI) : in broadleaf canopies, defined as the one-sided green leaf area per unit ground surface area
Fraction of Photosynthetically Active Radiation (FPAR) : a proportion of photosynthetically active light (400-700 nm) received by the canopy of plants
Land Surface Water Index (LSWI) : (Computed from Near-Infrared and Short-Wave Infrared bands) sensitive to total liquid water content in plant and soil backdrop [9]
These datasets were downloaded from the year 2001–2014. The spatial resolution is 500m and temporal resolution is 8 days.

2. Meteorological Features : The maximum temperature (Tmax), minimum temperature (Tmin), solar radiation (RAD), and vapour pressure (VHP9, VHP15; measured at 9:00 and 15:00, respectively) are retrieved from the Bureau of Meteorology, Government of Australia. The gridded product has a spatial resolution of 0.01 degree and a temporal resolution of 1 day. The data was downloaded from 2001 to 2014.

3. Topographical Features : Elevation and Latitude were also used as features.

The above data was collected for the 23 Flux sites in Australia as shown in Fig 1. GPP data measured at Flux Towers is taken as the ground truth. The dataset Fluxnet2015 contains information gathered from a number of regional flux network sites. It provides us the required ground truth. GPP from MODIS is also used for comparison against our model. To ensure homogeneity in terms of spatiotemporal resolution, the daily products are transformed into their 8-day average.



(a) The count of observations for each tower and year in the dataset

	count	mean	std	min	50%	max
EVI	4889.0	0.29	0.11	0.00	0.28	0.82
LSWI	4889.0	0.34	0.28	-0.35	0.32	0.94
FPAR	4889.0	0.49	0.23	0.02	0.46	1.00
LAI	4889.0	1.77	1.68	0.10	1.10	7.00
Tmax	4889.0	27.53	8.65	2.78	30.62	42.45
Tmin	4889.0	14.64	7.66	-3.75	15.11	26.86
VHP_9	4889.0	16.61	8.13	4.03	13.83	33.56
VHP_15	4889.0	14.87	7.31	3.96	12.48	32.74
RAD	4889.0	19.68	6.26	3.61	20.70	35.01
GPP	4889.0	4.10	3.51	0.00	3.38	25.31

(b) Data Summary for some features

Figure 2

2.3 Exploratory Analysis

The data contains 4889 observations with features *'Latitude', 'EVI', 'LSWI', 'FPAR', 'LAI', 'Tmax', 'Tmin', 'VHP9', 'VHP15', 'RAD', 'ELEVATION', 'GPP', 'DAYS', 'ID', 'YEAR' and 'Longitude'* as given in section 2.2. The *ID* represents the unique name of flux tower while *DAYS* and *YEAR* represents the Day and Year of the observation. Each Tower has atmost 46 observations (ie. DAYS numbered from 1 to 46) in a year, corresponding to an observation in every 8 days.

It should be noted from Fig 2(a), that the towers do not have observations over all the years. Also some years do not have all 46 observations. Hence, we can not treat this as a Time-Series problem. Instead we will treat each observation independent. This will allow us to easily fit any ML model utilizing the whole dataset.

Now, we shall observe the seasonal trend in GPP. As we have 23 towers, we will cluster the towers using only the tower locations (Latitude and Longitude). The clustering is done only location-wise as we expect the seasonal trends for towers close-by will be similar. The average GPP is then calculated within each cluster and plotted against days of observation. Using the k-Means algorithm, we obtain 4 optimal clusters as shown in Fig 3.

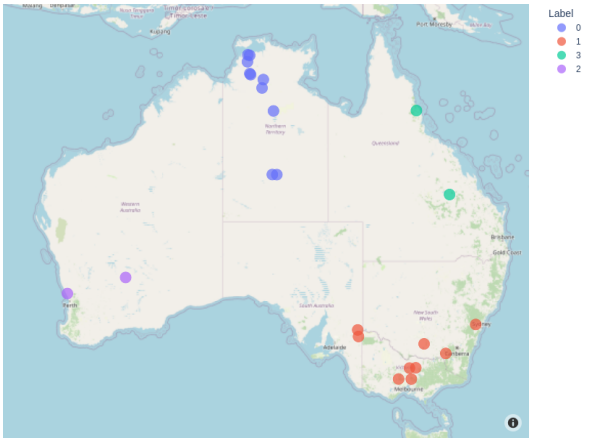
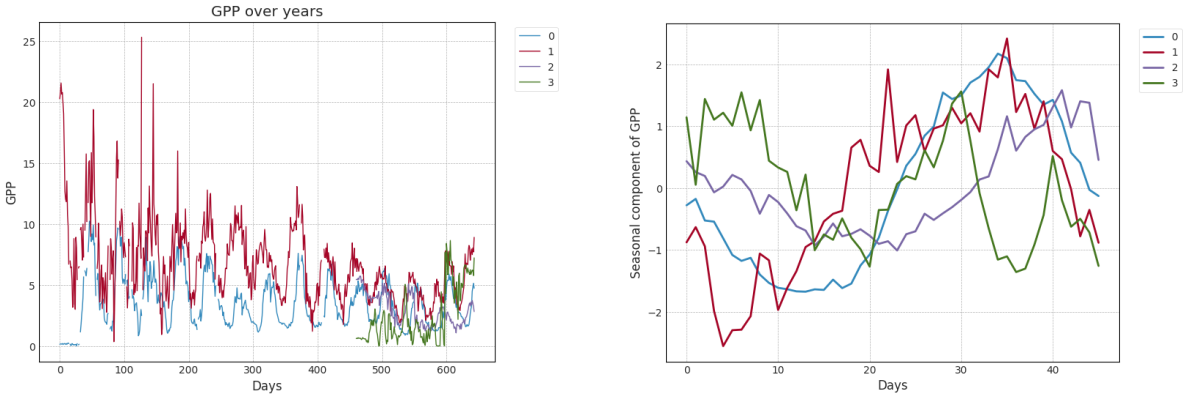


Figure 3: Towers divided into 4 clusters by k-Means

From Fig 4(a), there is clearly a seasonal trend in GPP. The seasonal component in Fig 4(b) is decomposed from the obtained Time Series of average GPP (considering days with observations from all towers within cluster) with



(a) GPP plotted over Time for the 4 clusters

(b) Seasonal Component of GPP for the 4 clusters

Figure 4

period 46. Seasonal Component shows GPP decreases in the beginning of the year, then increases and peaks around October after which it starts decreasing again. This can be explained by the Seasons cycle in Australia. September, October and November is the time of Spring where trees grow new leaves increasing the carbon flux. By the end of Summer in February and beginning of Autumn in March, April and May, trees start shedding leaves decreasing the GPP.

3 Modelling

We consider GPP obtained from Flux Towers as the ground truth. The features used to estimate it are *'Latitude', 'EVI', 'LSWI', 'FPAR', 'LAI', 'Tmax', 'Tmin', 'VHP9', 'VHP15', 'RAD' and 'ELEVATION'*. The satellite-based observations are re-sampled to 500 m spatial resolution by bilinear interpolation method and the daily observations are converted to their 8 days average ensuring homogeneity in terms of spatiotemporal resolution.

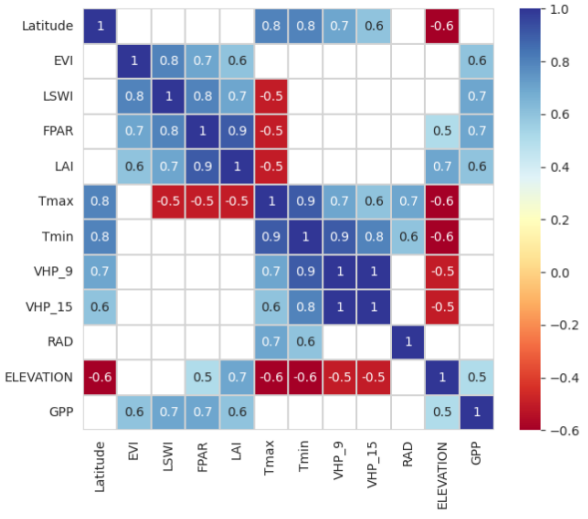


Figure 5: Correlation matrix masked to show absolute values greater than 0.5

The Pearson correlation matrix in Fig 5 shows presence of linear correlation between GPP and EVI, LSWI, FPAR and LAI. We also notice presence of strong multi-collinearity within features such as :

(i) EVI, LSWI, FPAR, LAI : all these features are high in a healthy ecosystem and vice versa.

(ii) Tmax, Tmin, VHP9, VHP15 : all these features are high for high temperatures and vice versa.

Multicollinearity will be an issue in inference of the linear regression model, but will not be a problem for prediction and for other non-linear models we will consider like Random Forest, SVM.

We will consider the data for years 2001 to 2013 as training set and the year 2014 as our testing dataset since then sufficient data is available to evaluate our model.

3.1 Fitting Multiple Linear Regression Model

Fitting a Multiple Linear Regression Model using the above mentioned features and Fluxnet GPP as response gives us training **R-squared of 67%** and **Adj. R-squared of 66.9%**.

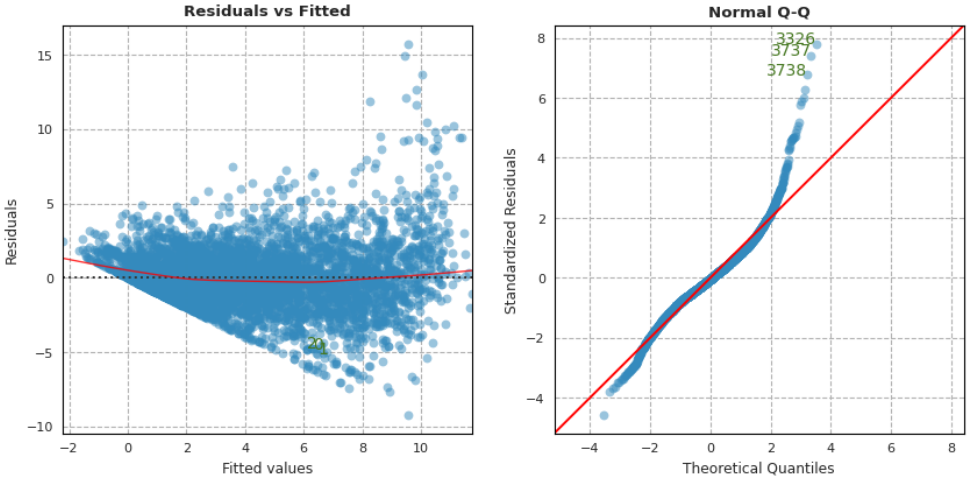


Figure 6: Residual Plots for the LR model

In Fig 6, the QQ plot suggests that the residuals do not follow Normal distribution, also suggested by Shapiro Wilk test which gives $<<$ 0.05 p-value. The Residuals vs Fitted plot suggests presence of heteroscedasticity in our model (Breush Pegan test gives $<<$ 0.05 p-value). There was no presence of influential points in our data. The Durbin Watson test statistic (=0.530) is very close to 0 suggesting presence of positive auto correlation. This added with presence of strong multicollinearity defies all assumptions of Linear Regression essential for inference. Studying the Partial Residual Plots also suggests presence of non linearity in some features, especially *Tmax*.

To create a better model, we will use the squared term of *Tmax* and use Box-Cox Transformation on the response. The optimal parameter λ found by maximizing the maximum log likelihood for different λ's is **0.231**.

$$y \longleftarrow \frac{(y+1)^{\lambda}-1}{\lambda} \tag{1}$$

Fitting a Multiple Linear Regression Model with the above changes gives us training **R-squared of 70.1%** and **Adj. R-squared of 70%**. The test R-squared for year 2014 is **71.46%**. The test **RMSE = 1.8623** and **MAE = 1.2896** (obtained after inverting the response to original form).

3.2 Fitting Polynomial Regression Model

For every feature, we now consider its 2nd and 3rd exponents as additional covariates. Fitting a Multiple Linear Regression Model on this data gives us training **R-squared = 74.6%, Adj. R-squared = 74.4%**. As Adj. R-squared is still close to R-squared, we did not add many meaningless variables.

But, like the linear regression models, the model assumptions of homoscedasticity, normality in residuals, multicollinearity are not satisfied. Adding polynomial terms improved the performance but did not help in making the model available for inference, hence we should move away from the linear models setup and try some non linear models.

3.3 Fitting Non-Linear Regression Models

Now, we will try various non-linear models as the assumptions of linear regression including that of linearity with response were not satisfied. We will be fitting Support Vector Machine (SVM), Tree based model Regression Tree and Ensembled model Random Forest.

3.3.1 Support Vector Regression

Support Vector Regression [10] works on the same principles as SVMs. The best fit line in SVR is the hyperplane with the greatest number of points. In contrast to other Regression models, which strive to reduce the error between the real and projected values, the SVR seeks to match the best line within a threshold value. The threshold value is the distance between the boundary line and the hyperplane.

Without raising the computing cost, a kernel aids in the discovery of a hyperplane in the higher dimensional space which is essential if we cannot locate a separating hyperplane in a particular dimension. We considered popular kernels like **Linear, Polynomial** *(degree = 3)*, **Radial basis function (RBF)** for the problem.

Fitting SVR using 5-fold cross validation, the Mean RMSE and MAE are given in Fig 7. The RBF kernel gives us better results against the rest. The RBF kernel on two samples $\mathbf{x} \in R^k$ and **x'**, is defined as

	RMSE	MAE
rbf	2.602032	2.024879
poly	2.875822	2.272227
linear	2.892434	2.282381

Figure 7: Mean R-squared and RMSE for different kernels

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\sigma^2}\right) \tag{2}$$

Yet the RMSE value are very high, suggesting that neither kernel is able to capture the non linearity present in our data. Hence, we shall move to other models.

3.3.2 Regression Tree

A regression tree iteratively splits the space into sub-spaces based on a splitting criteria until a stopping criteria is achieved. The splitting criteria used is minimizing the Mean Squared Error. Tuning the parameter max_depth i.e. maximum allowed depth, using 10 fold cross-validation repeated 3 times gives us an optimal value of **5**. The Mean RMSE and Mean MAE are **2.1373** and **1.6026** respectively.

max_features	3	4	5	6	7	8
RMSE	1.729 (0.129)	1.727 (0.129)	1.725 (0.130)	1.726 (0.129)	1.727 (0.130)	1.729 (0.128)

Table 1: Mean RMSE with std. deviation for *max_features* ranging from 3 to 9

3.3.3 Random Forest Regression

Random Forest Regression [11] creates an ensemble of Regression Trees where a number of trees (*n_estimators*) are each trained on a bootstrap sample of the training dataset with a random selection of features. We consider the forest with 1500 trees and tune the model to find optimal value of hyper-parameter *max_features* which is the number of features used to find best split. The max depth for each tree is set at 8 and the measure of quality of a split is mean squared error.

Using 10 fold cross-validation repeated 3 times, we obtain the following RMSE values for different *max_features* (Total number of features being 11).

As the Mean RMSE is lowest for *max_features* = 5, we will choose it as the optimal value. In Fig 8, the Training and Test RMSE is plotted for the RF model with *max_features* = 5 and different values of *max_depth*. To prevent over-fitting over the training data, we choose the optimal value of *max_depth* as 8. Since it ensures the train and test RMSE are sufficiently low and the difference between them is not large indicating over-fitting.

Fitting a Random Forest with the tuned hyper-parameters gives us training **R-squared of 85.70%**, **RMSE = 1.2660** and **MAE = 0.8679**.

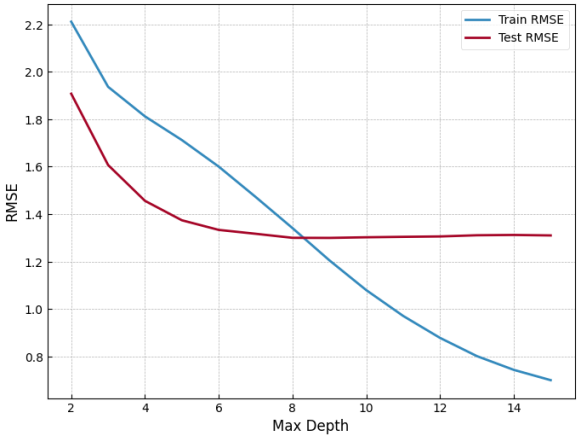


Figure 8: Train and Test RMSE for max_depth ranging from 2 to 15

4 Model Evaluation

We choose the tuned Random Forest as our final model as it gives us the highest training R-squared among the rest. Predicting the GPP values for the test data *(year 2014)*, we obtain the **Test R-squared is 83.06%, RMSE = 1.3062, MAE = 0.8955**.

Fig 9 shows the impurity-based feature importance of the 11 features. We notice *LAI* and *LSWI* are the most important features followed by *EVI*, *Elevation*, etc. LAI estimates the amount of leaf material in a canopy, which directly determines the amount of photosynthesis achievable. The Land Surface Water Index quantifies the increase in liquid water content in soil and vegetation, which influences the growth and health of trees in the region [9]. So it is justified that these are the most important features found by the model.

	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
Performance	4.125107	2.808718	3.14257	2.313214	1.571569	1.434136	2.418737	1.890141	1.985651	1.400974	1.546903	1.779904	1.340068	1.300145

Figure 10: RMSE values for leave one year out

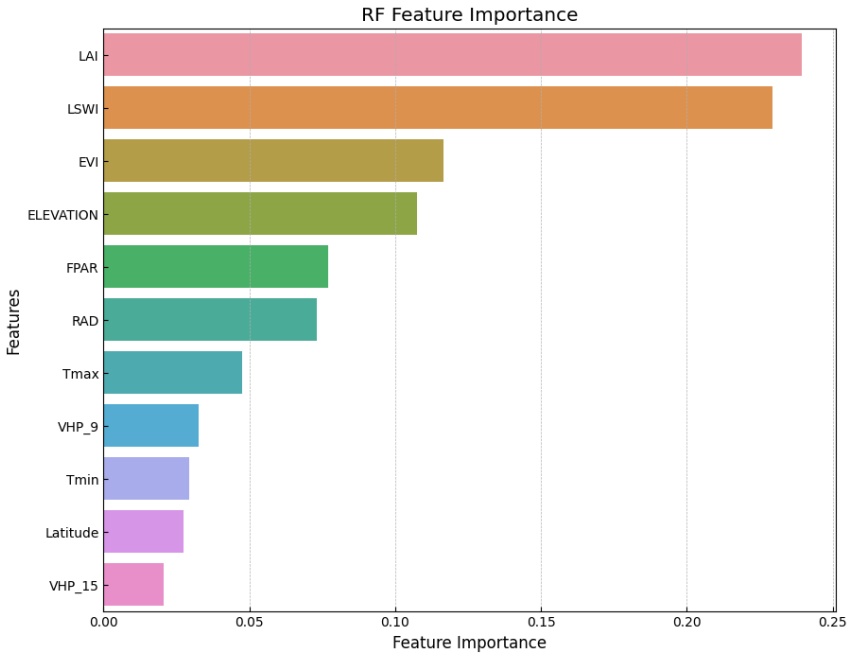


Figure 9: Feature Importance for RF model

The Mean Decrease in Impurity (MDI) is a metric used in to measure the importance of features. It quantifies the average amount by which the impurity measure (Entropy in case of Regression) decreases when a particular feature is used for splitting in the decision tree.

The MODIS GPP product also provides us with the GPP estimates at a global scale. The estimates at the 23 Flux sites in Australia were downloaded with 500 m spatial resolution and 8 days temporal resolution. The R-squared between MODIS GPP and FLUXNET GPP over the years 2002 to 2014 is 44.39%, while that for year 2014 is 36.41%. Hence our model outperforms the MODIS GPP estimates and can upscale GPP from towers to regional scope.

To judge the performance of model over the years and across all Flux Towers, we find the RMSE Leaving One Year Out and Leaving One Tower Out as shown in Fig 10 and Fig 11. In Leaving one year out, the models are trained on data leaving one year out and then tested on the left out year (Similarly Leave one Tower out). The RMSE values suggests that the model performs better for the later years (after 2010). This can be attributed to presence of very few data from towers in earlier years (Fig 2) and upgrade of sensors in the Flux Towers making the measurements more accurate and less susceptible to errors in later years.

On the other hand, RMSE values for leaving one tower out suggests

	Performance
FLX_AU-Ade	2.148476
FLX_AU-ASM	0.933776
FLX_AU-Cpr	0.56697
FLX_AU-Cum	1.180422
FLX_AU-DaP	2.692738
FLX_AU-DaS	1.74499
FLX_AU-Dry	1.222292
FLX_AU-Emr	1.403384
FLX_AU-Fog	1.921897
FLX_AU-Gin	1.688583
FLX_AU-GWW	0.803527
FLX_AU-How	2.130051
FLX_AU-Lox	3.707273
FLX_AU-RDF	1.799397
FLX_AU-Rig	1.63767
FLX_AU-Stp	1.209372
FLX_AU-TTE	0.508964
FLX_AU-Tum	4.890827
FLX_AU-Wac	1.675555
FLX_AU-Whr	1.606567
FLX_AU-Wom	1.564876
FLX_AU-Ync	1.174744
FLX_AU-Rob	1.943935

Figure 11: RMSE values for leave one tower out

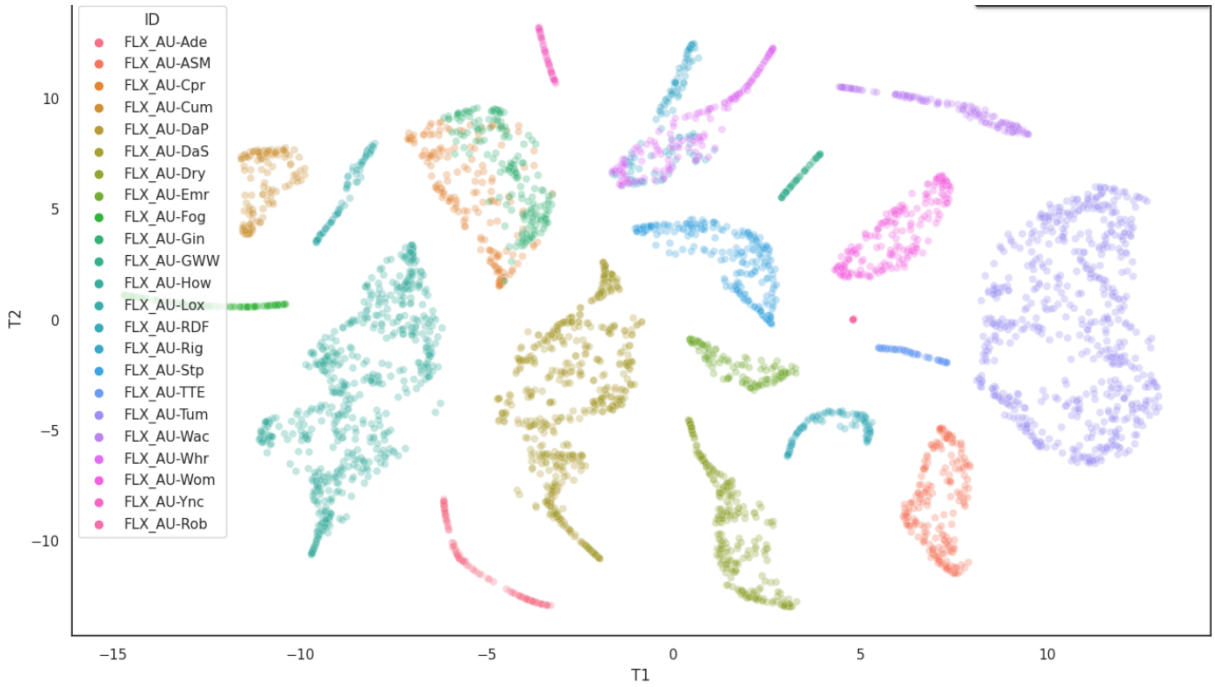


Figure 12: First 2 components of t-SNE for Australian data

that some towers are worse to predict like *FLX_AU-Tum, FLX_AU-DaP*, etc than the rest. To better inspect this phenomenon we will use t-SNE to visualize our high-dimensional data in a 2-dimensional space.

t-distributed stochastic neighbor embedding (t-SNE) is a non-linear dimensionality reduction technique which embeds the data into lower dimension, meaning if the points were close in higher dimension then they will be close in the embedding, and vice versa, with a high probability.

Fig 12 shows the t-SNE visualization (2 components) of the Australian dataset and we can see that the the observations from the same tower are similar to each other. While, observations from different towers are highly dissimilar, barring a few towers like *FLX_AU-Gin, FLX_AU-Cpr*, etc. We notice that the more dissimilar, observations of a tower are from the rest, the higher is their corresponding Leave one Tower out RMSE. Which is justified as no other tower's observations can explain them, increasing the prediction error.

5 Extending to Indian terrain

Now we shall expand our work to Indian terrain which has only 4 Flux towers situated in Haldwani (Uttarakhand), Barkot (Uttarakhand), Meerut (Uttar Pradesh) and Betul (Madhya Pradesh), with very low data availability. Since this wont be enough to train a good model which works all over India, we will follow a different approach.

We will instead use a Global dataset for training the model, so that we have enough data for training. A good model will learn the variation in GPP over different terrains which might help it predict GPP for India reasonably. Fig 13 shows location of the 59 Flux towers present in Evergreen regions considered for training. The data ranges from 2001 to 2014 and the observations are recorded monthly (Just like 2, this data also has missing years and some missing observations). The total number observations is 5029 with 20 features and GPP.

Figure 13: Map of the 59 global Flux tower sites in Evergreen regions

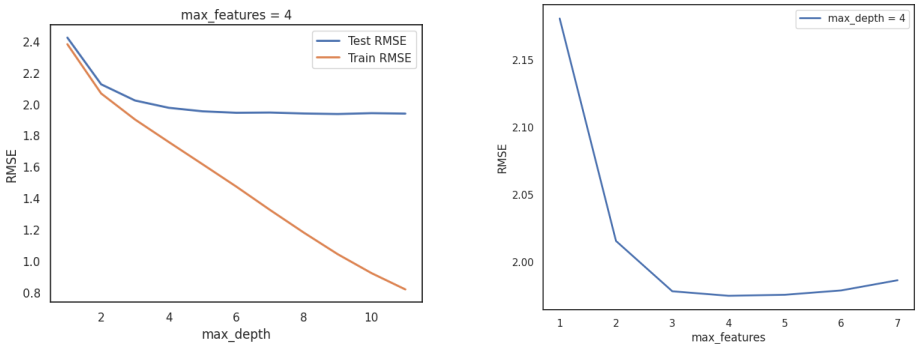
5.1 Fitting RF Model

Next, we train Random Forest models on the training dataset with 1000 trees and varying set of hyper-parameters *max_depth* and *max_features*. The test and train RMSE(and MAE) values for each model are computed with 5-fold cross validation. To prevent overfitting we choose a depth for which the training RMSE is not considerably lower than the test RMSE. This gives us optimal $max\_depth = 4$ (Fig 14(a)). Fig 14(b) shows the lowest RMSE is achieved when $max\_features = 4$. So, 4 is the optimal value of *max_features*. The same set of optimal values was found through MAE as well.

The Training performance metrics for this model are as follows : **RMSE = 1.7960** , **MAE = 1.2092** , **R squared = 0.7503**.

5.2 Prediction of RF Model

Using the RF model trained on global dataset, we will now predict GPP over the Indian region. For this purpose, we considered the monthly satellite data of the features over the evergreen



(a) RMSE of the RF model plotted against max_depth

(b) RMSE of the RF model plotted against max_features

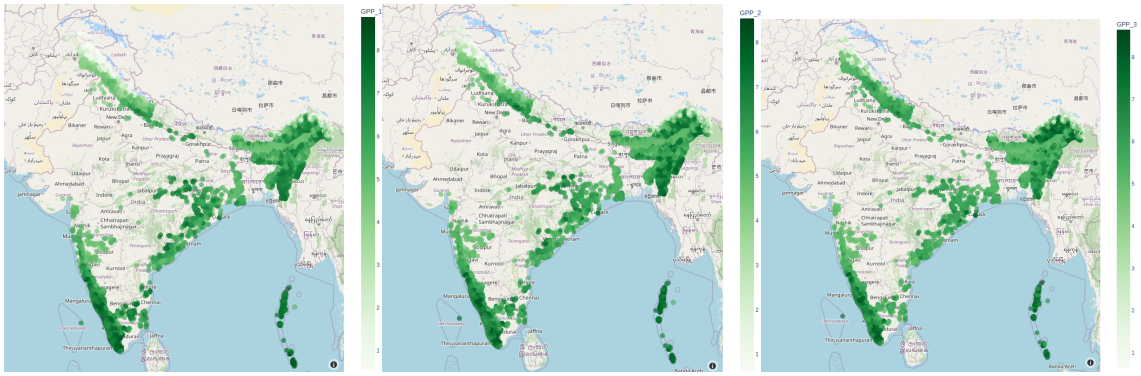
Figure 14

forests of India between years 2001-2020. Only evergreen forests were considered as the model is trained only at evergreen sites, hence it will minimize the unknown variability. The locations corresponding to Evergreen forests were identified using Land Use/Land Cover (LULC) maps. The satellite data is spatially dense, having resolution of $1km^2$, and contains data for *424,491* locations over the Indian terrain. For each of these locations, we have 240 observations corresponding to a monthly observation over a period of 20 years (2001-20).

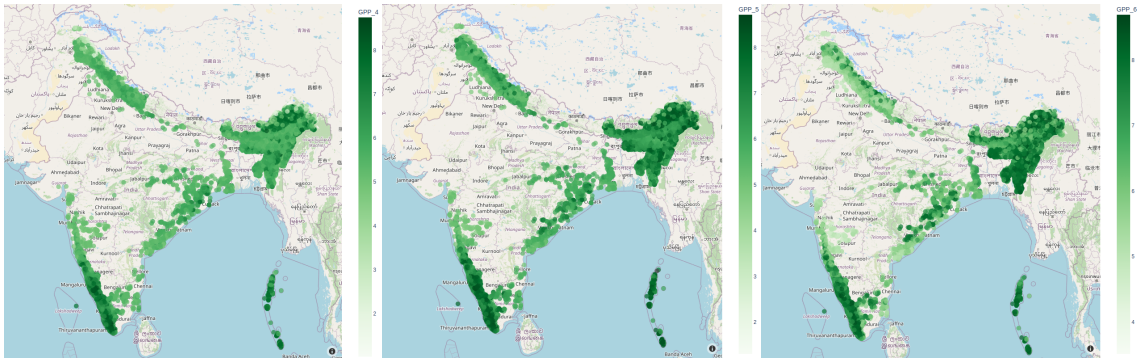
The predictions for the year 2014 are shown in Fig 15. The green coloured regions correspond to the satellite data of evergreen forests and the intensity of green colour corresponds to the predicted GPP (higher prediction is shown with darker green). We can observe the expected patterns in monthly plots. The GPP overall is relatively lower in December, January, February, March due to winter season. Lower temperatures and Solar Radiation causes lower GPP values. This is most significant in high elevation areas like Jammu and Kashmir and areas near Nepal border, as they are further impacted by snow. After winter GPP values slowly increases with Spring season and reaches its peak by end of summer (June). As Autumn season (October, November) enters, the GPP values start to decrease. During Spring, trees create new leaves increasing their LAI (Leaf Area Index), increasing GPP. While during Autumn, trees shed leaves decreasing the LAI. Hence GPP decreases.

The seasonality can be expressed better in Fig 16(a), which is a trend plot of predicted GPP between 2011-20, averaged over all available evergreen locations in the Indian dataset. We can clearly observe the expected seasonality in the predicted GPP. The GPP is highest in July, i.e. the end of summer, after which it starts decreasing in Monsoon and Autumn. The GPP is lowest during winters and then increases in Spring and Summer, An another interesting observation is that the peaks of the predicted GPP are slowly shifting forwards, which is inline with the late monsoons due to climate change.

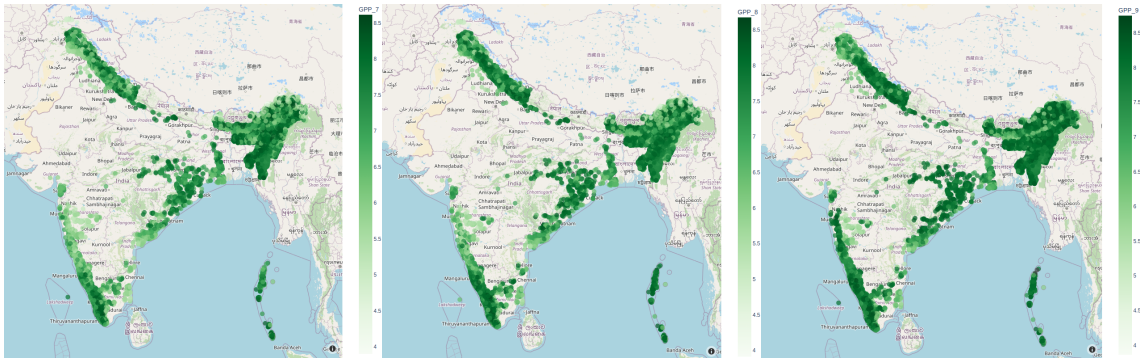
Next, we observe the rate of change in predicted GPP over every location between 2011-20 in Fig 16(b). This problem is same as finding the slope of predicted GPP against time between 2011-20, at every location. We use the Theil-Sen estimator [12], [13] to find this slope instead of the Linear Regression estimate of slope. The Theil-Sen estimate is a non-parametric estimate for the slope and is more robust than the OLS estimate in case of outliers and for skewed, heteroskedastic and non-normal data. It is calculated by finding the median of slope of all lines passing through any pair of points. The Kendall-tau statistic test between the OLS slope and Theil-Sen slope is **0.4785** with p-value $<<$ 0.05. This signifies that the two kinds of slope are indeed correlated but the Theil-Sen estimator is more robust for non-normal dataset, as in our case. Hence, we use the Theil-Sen estimates instead.



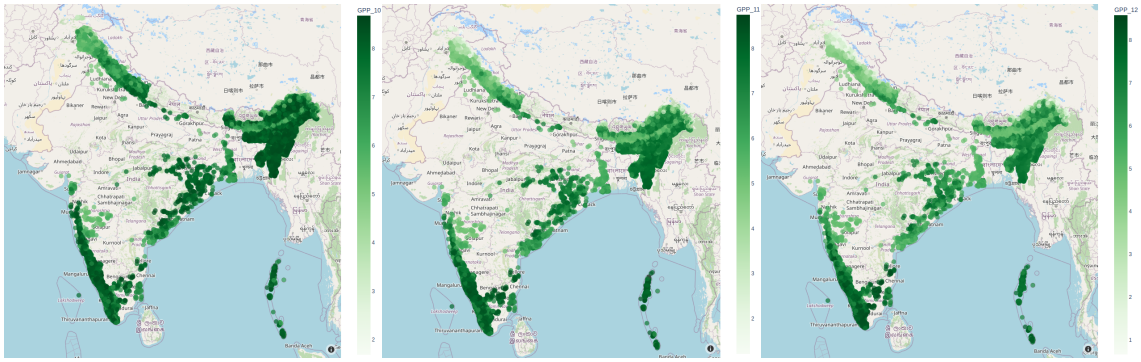
(a) January (b) February (c) March



(d) April (e) May (f) June

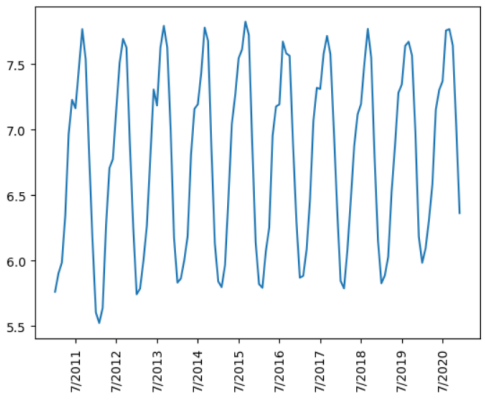


(g) July (h) August (i) September

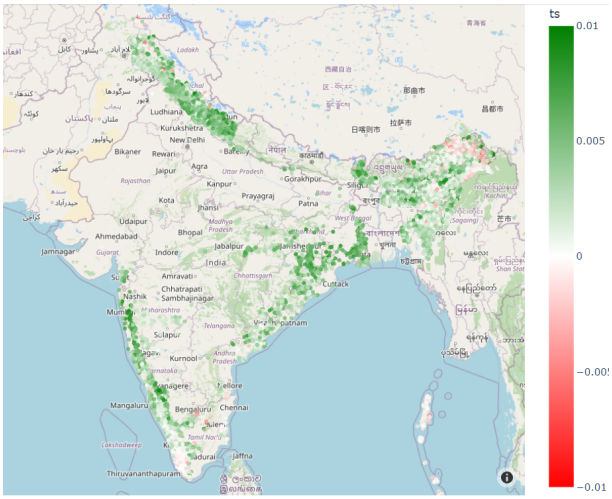


(j) October (k) November (l) December

Figure 15: GPP estimated for every month of year 2014



(a) Average Predicted GPP over the Indian terrain from 2011-20



(b) Rate of change in Predicted GPP over the Indian terrain from 2011-20, by Theil-Sen estimator (the color-bar ranges between -0.01 to 0.01)

Figure 16

The data for 3 Indian Flux towers located in Haldwani, Barkot and Meerut was used to validate the model. This is monthly data available between 2016 to 2018 for Haldwani and Meerut, and only 2016 for Barkot.

The test RMSE obtained over the data of 3 Indian Flux Towers is **1.892**. But this validates our model only at 3 specific locations and not over all the prediction locations.

6 Estimating Performance without ground truth

The predicted GPP obtained by training on the Global dataset seems to follow the expected pattern but we want to know if it is truly close to the real GPP values. If it is actually close, we were successful in capturing variation of GPP in evergreen regions of India by training on evergreen regions of other places combined.

But to check the performance of our model predictions we will need to know the true GPP values at all the *424,491* locations over the Indian terrain. Since, we do not have the true GPP values, we need to estimate the model's performance when ground truth (i.e. response) is not available.

This raises two important questions that we need to answer.

- Is there a concept drift when considering Indian dataset from the global dataset ?
- How to estimate the performance measure (like RMSE, MAE) for Indian dataset ?

6.1 Drifter Algorithm

Concept drift occurs when the distribution of data changes over time. Presence of concept drift in the test-train data can lead to large errors in regression estimates. Real Concept drift is the

change in conditional probability distribution $p(y|x)$ and Virtual Concept drift refers to change in the distribution of covariates $p(x)$. Without the presence of ground truth, we can only detect presence of virtual concept drift.

The paper [14] suggests solving the following problem. For a regression function f and a threshold σ. Predict if generalization error E of f on testing data satisfies $E \geq \sigma$ when response on test data is unknown.

The threshold σ is chosen such that, if $E \geq \sigma$ we consider it as presence of virtual concept drift. The paper aims to find a measure d that is monotonic with the true RMSE. Hence, $E \geq \sigma$ implies $d \geq \delta$ for some optimal value of δ.

Let $D_{tr} = \{(i, x_i, y_i)\}_{i=1}^{n_{tr}}$ be the training data, $D_{te} = \{(i, x'_i, y'_i)\}_{i=1}^{n_{te}}$ be the Test data and $D_{|s} = \{(i, x'_i, y'_i) | a \leq i \leq b\}$ be a segment of the data are defined by tuples $s = (a, b)$ where $a \leq b$. Then for a regression function $f : R^m \to R$ trained using D_{tr} , The generalization error of f on the data set $D = \{(i, x_i, y_i)\}_{i=1}^{n}$ is defined as

$$RMSE(f, D) = \left(\frac{1}{n} \sum_{i=1}^{n} [f(x_i) - y_i]^2 \right)^{1/2}$$

The main idea, is instead of using $[f(x_i) - y_i]^2$ as the distance measure we use

$$d(x_i) = [f(x_i) - f'(x_i)]^2$$

where f and f' are two regression functions trained on different subsets of the training data. This way less important feature will have a low effect on the distance measure and it still measures how far x_i is from D_{tr}, small values of $d(x_i)$ suggests we are close to training set and that the prediction $f(x_i)$ is a good estimate. So large values of $d(x_i)$ indicate presence of virtual concept drift.

Proof for Linear Models :

For a linear model under usual notations, $\mathbf{y} = \mathbf{X}\beta + \epsilon$; where $\beta \in R^m$, $X_{n \times m}$ is the data matrix and ϵ_i are independent random variables with zero mean and variance of σ_y^2. Consider 2 OLS models trained :

- $\hat{y} = f(x) = \hat{\beta}^T x$ trained on dataset of size n
- $\hat{y'} = f'(x) = \hat{\beta'}^T x$ trained on independently sampled dataset of size n'.

Then the MSE $E[(f(x) - y)^2]$ is monotonically related with the expected squared distance measure between f and f'.

$$E[(f(x) - y)^2] = (1 + n/n')^{-1} E[(f(x) - f'(x))^2] + \sigma_y^2$$

Proof

Assume we have centered covariates such that $x_{i,1} = 1$ and $E[x_{i,j}] = 0$ for all $j \neq 1$, where $j = 1$ corresponds to the intercept. Also we assume that the axes of covariates have been rotated such that the covariates are uncorrelated. $E[x_{i,j} x_{i,k}] = \sigma_{j,k}^2 \delta_{j,k}$, where $\delta_{j,k} = 1$ if $j = k$ and 0 otherwise

For a dataset of size n, the OLS estimate of β, denoted by $\hat{\beta}$, is a random variable that obeys a normal distribution with a mean of β and a covariance given by $n^{-1}\Sigma$

$$Var(\hat{\beta}) = \sigma_y^2(X^TX)^{-1} = \sigma_y^2(n\ diag\{1, \sigma_{2,2}^2, ..., \sigma_{m,m}^2\})^{-1} = n^{-1}\Sigma$$

where $\Sigma = \sigma_y^2 diag\{1, \sigma_{2,2}^{-2}, ..., \sigma_{m,m}^{-2}\}$.

Hence,

$$E[(f(x) - y)^2] = Var(x^T\hat{\beta} - y) = x^T(n^{-1}\Sigma)x + \sigma_y^2$$

and

$$E[(f(x) - f'(x))^2] = Var(x^T\hat{\beta} - x^T\hat{\beta}') = x^T(n^{-1}\Sigma)x + x^T(n'^{-1}\Sigma)x = x^T(n^{-1} + n'^{-1}\Sigma)x$$

Replacing $x^T\Sigma x$ in first equation we get ,

$$E[(f(x) - y)^2] = (1 + n/n')^{-1}E[(f(x) - f'(x))^2] + \sigma_y^2$$

Hence proved.

For non linear models, such a proof is difficult but the results have been shown to hold in [14] through simulation studies.

To apply Drifter algorithm we go through the following steps :

6.1.1 Training phase

- Train the main model f on D_{tr}.
- Consider l subsequences $(s_1, s_2, ..., s_l)$ in $[n_{tr}]$. These are called "segments" of training data. The segments are chosen such that the data in each segment corresponds to only one "concept". Segments can be overlapping and of different lengths to make the process more robust. Segmentation consisting of equally-sized segments of length l_{tr} with 50% overlap is quite robust.
- Train l segment models, i.e. model f_i is trained on $D_{|s_i}$, $i = 1(1)l$. The segment models be of different/weaker form than the main model f.

In our case performing *t-SNE* on the global dataset shows that towers are majorly separable and each tower explains a different region of variation. Hence we choose to divide our data into 59 segments where each segment corresponding to a different tower. We train 59 Random Forest segment models and obtained an average RMSE = 0.6101, MAE = 0.44 and R squared = 0.9126. This suggests we have segment models that can explain their concept well.

6.1.2 Testing phase

- For each of the segment model, find RMSE on D_{te}

$$z_i = RMSE(f, f_i, D_{te}) = \left(\frac{1}{n_{te}}\sum_{j=1}^{n_{te}}[f(x'_j) - f_i(x'_j)]^2\right)^{1/2}$$

- This provides us l estimates of the generalisation error z_i, and we choose the n_{ind} 'th least value as the value for the concept drift indicator variable d.
- For an overlapping segmentation technique, $n_{ind} = 2$ can be employed since it is plausible to expect that at least two of the segment models should have small z_i values if the testing data has no concept drift, but a single small z_i value could still occur by chance.

We choose the smallest z_i as the concept drift indicator variable d as we don't have overlapping segments. The value obtained is $\mathbf{d = 0.70828}$.

6.1.3 Choosing Threshold

We first need to fix σ i.e. the threshold for true generalization error for concept drift. As the segments we chose explains one unique concept (observed by performing *t-SNE*) in the data, each segment has a concept drift against the remaining segments.

Let a regression model f be trained on all segments
For $i \in \{1, 2, ..., 59\}$,
Let a regression model f_i be trained on all segments except i'th segment
Let $\sigma_i = RMSE(f_i, D_{|s_i})$,i.e. RMSE when tested on i'th segment
Let d_i be the concept drift indicator variable calculated (Section 6.1.1) for the i'th segment

Then we choose σ as:

$$\sigma = \left(\frac{1}{n_{tr}} \sum_{j=1}^{n_{tr}} [y_j - \hat{y_j}]^2 \right)^{1/2}$$

where $\hat{y_j}$ is predicted using f_i if y_j is part of i'th segment. As this is the error that can be explained by the model f, since all these unique segments are part of it's training data. For our data we get $\sigma = \mathbf{2.082}$, i.e. if true RMSE is greater that 2.082 we shall call it a virtual concept drift.

Here we can also show that $\sigma_i's$ and $d_i's$ do indeed have a monotonic behaviour. To test this we use the two-sided non-parametric kendall's tau statistic test to test H_0 : if the two variables are uncorrelated. We obtain kendall's tau = 0.680 and corresponding p-value ≤ 0.05, suggesting presence of correlation among them.

Now, for choosing threshold for d , i.e. δ , we define the following setup:
Let $\delta > 0$, then,

.	$d_i < \delta$	$d_i \geq \delta$
$\sigma_i < \sigma$	TP	FN
$\sigma_i \geq \sigma$	FP	TN

Table 2: Setup to find optimal δ

Here TP (True Positives) = number of segments with $\sigma_i < \sigma$ (segment is not concept drifting) and $d_i < \delta$ (δ is able to classify the segment to non concept drifting).

Essentially, we have two classes {True = Concept drift Absent in segment, False = Concept drift Present in segment }. Hence a segment i belongs to class 'True' if $\sigma_i < \sigma$ and vice versa. We consider a classifier $\{d_i < \delta; \delta > 0\}$ to classify the segments. Now, our goal reduces to finding a δ which results in the best classifier.

We can do this by finding δ that maximizes the $F1$ score of classifier, which is a metric that combines precision and recall of the model into a single score, providing a balanced measure of the model's effectiveness. A high F1 score indicates that the model has achieved a good balance between precision and recall. Precision measures the accuracy of the positive predictions made by the model, while recall measures the model's ability to correctly identify all positive instances in the dataset.

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall}$$

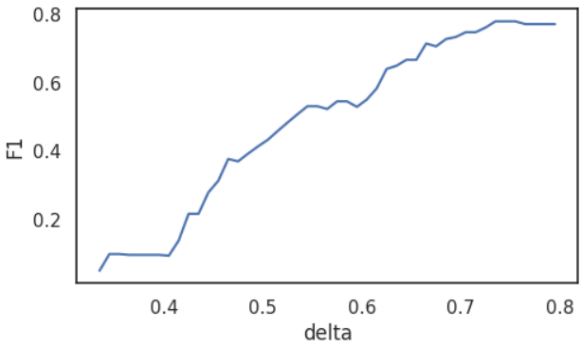


Figure 17: F1 score versus δ

Fig 17 shows the maximum F1 score is achieved at $\delta = \mathbf{0.7551}$.

In 6.1.1 we found $d = 0.70828$. Hence, the Indian Dataset used for testing does not have a significant concept drift ($d < \delta$) that cannot be explained by other towers combined in the global dataset. However the value isn't is comparable to the threshold value, which is expected as the Indian dataset is still located far away from the other towers in global dataset, making it difficult to estimate GPP for it through data from other towers.

6.2 Direct Loss Estimation (DLE) [15]

Now that we know that, there is no concept drift in Indian dataset against the global dataset. We can train the DLE algorithm on global dataset and use it on Indian dataset to find the estimated Loss.

Here the data is divided into 3 categories :

- Training Set : (X_{train}, Y_{train})
- Reference Set : (X_{ref}, Y_{ref})
- Testing/Analysis Set : (X_{test})

And two models, *child* and *nanny* are trained as follows :

6.2.1 Training Phase

- Train the Child model on Training set

$$X_{train} \rightarrow Child \rightarrow Y_{train}$$

- Predict Reference set using the Child model

$$X_{ref} \rightarrow Child \rightarrow \hat{Y}_{ref}$$

- To train the Nanny model, $\hat{Y}_{ref}$ is also included as a covariate in X_ref. The response is the Loss done by child model on reference set, i.e the j'th response is taken as $|Y_{ref,j} - \hat{Y}_{ref,j}|$

$$(X_{ref}, \hat{Y}_{ref} \rightarrow Nanny \rightarrow Loss(Y_{ref}, \hat{Y}_{ref})$$

In our dataset, to maintain homogeneity in train and reference dataset, we did stratified sampling to divide the whole global dataset into train and reference of equal sizes. We used Random forest for both child and nanny models and trained them as mentioned. On the reference set, the child model gives an RMSE of 1.7939. The Nanny model has the following training performance metrics : **RMSE = 0.4348, MAE = 0.2396, R squared = 0.8903** Hence we have trained reasonably good child and nanny model.

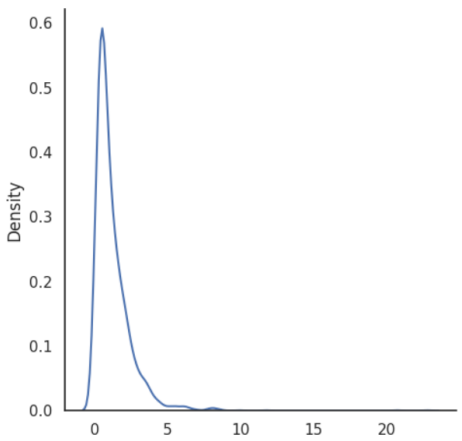


Figure 18: Density of absolute error of Child model on Reference set

6.2.2 Testing Phase

- Predict the Child model on Testing set

$$X_{test} \rightarrow Child \rightarrow \hat{Y}_{test}$$

- To get Loss estimate, predict Nanny model as

$$(X_{test}, \hat{Y}_{test} \rightarrow Nanny \rightarrow \hat{Loss}(Y_{test}, \hat{Y}_{test})$$

Testing the models as above, gives us a vector of Loss estimates for each observation in the test dataset. This can then be manipulated to find estimates of RMSE and MAE.

$$RMSE = \left(\frac{1}{n_{te}} \sum_{j=1}^{n_{te}} \hat{Loss}_j^2(Y_{test}, \hat{Y}_{test}) \right)^{1/2} = 1.9370$$

$$MAE = \frac{1}{n_{te}} \sum_{j=1}^{n_{te}} \hat{Loss}_j(Y_{test}, \hat{Y}_{test}) = 1.8863$$

Hence, using the Random Forest model we get estimated $RMSE = 1.9370$ and $MAE = 1.8863$ metrics for test data. Comparing this to SVM Model, which has $RMSE = 2.85$ and $MAE = 2.64$, gives us that the Random Forest model works much better than other models for prediction.

7 Future Work

Other methods for evaluating model in absence of ground truth can be explored. Bayesian Approach gives us a probabilistic prediction for the GPP, this can be of great use with many interesting possibilities. But challenges in this approach are choosing a suitable prior of the parameters of the model and it is very computationally expensive, especially for a large dataset like in our case.

8 Discussion / Results

The important results we found have been listed below :

- Over the Australian region, we found that our tuned Random Forest model performs the best with training R-squared of 85.70%, RMSE = 1.2660 and MAE = 0.8679.
- LAI and LSWI are the most important features for predicting GPP followed by EVI, Elevation, etc.
- Over the Indian region, our tuned Random Forest model performed the best with training RMSE = 1.7960 , MAE = 1.2092 and R squared = 0.7503. Testing it over the available data of 3 Indian Flux towers gives RMSE = 1.892. This suggests that over the true GPP available for validating the model, our model does indeed perform very well. The predicted GPP also follows the expected trends, further instilling confidence in the model.
- Since we do not have true GPP values for the whole testing region, we used DLE algorithm to estimate the RMSE and MAE values. But the algorithm assumes that concept drift is absent in the data. So, we used the Drifter algorithm to check the presence of concept drift with the above RF model as base model. Using the towers in the global training data as the different segments, we get concept drift indicator variable $d = 0.708$. Using the procedure in 6.1.2, the optimal threshold for d is $\delta = 0.755$. As $d < \delta$, we conclude that there is no concept drift in Indian dataset against global dataset.
- Now, the DLE algorithm gives the estimated RMSE = 1.9370 and MAE = 1.8863 for the RF model over test data. This suggests our model does indeed predict the GPP values over Indian terrain well and better when compared with other models like SVM.

9 Conclusion

The Gross Primary Productivity (GPP) is one of the very important measures to know the health of an ecosystem. In this work, we wanted to explore estimating GPP using multisource data (i.e. including Biophysical, Meteorological and Topographical features). We first did it for Australia, which has plenty of data for both training and testing. This concluded that we can indeed estimate GPP with the data and achieve much better performance than MODIS estimates. We also found that LAI and LSWI are the most important features followed by EVI, Elevation, etc. The next problem we tackled, was estimating GPP in Indian terrain where we do not have enough data to even train. We found that using global dataset in such situations can still deliver good performance and that a new 'concept' introduced by this region is already explained in the global dataset. Even after training on global dataset, we were able to achieve the expected predictions on Indian Dataset using the Random Forest model. The predicted values have the seasonality and distribution like we expected. Using the Drifter algorithm we found that this model is indeed able to explain the variation of GPP in Indian region. This suggests that even though we didn't use Indian data for training, the model was able to use data from other similar global towers to predict GPP over this region without a significant error. Using DLE algorithm, we found the estimates of RMSE and MAE for the RF model over Indian test data and saw that our model performed the best compared to others.

Acknowledgement

I would like to express my gratitude to Mr. Deep Prakash Sarkar for providing me with the data of Australian and Indian regions, and to Dr. B. Uma Shankar for his continuous guidance and motivation.

References

[1] C. Beer, P. Ciais, M. Reichstein, *et al.*, "Temporal and among-site variability of inherent water use efficiency at the ecosystem level," *Global biogeochemical cycles*, vol. 23, no. 2, 2009.

[2] J. L. Monteith, "Solar radiation and productivity in tropical ecosystems," *Journal of applied ecology*, vol. 9, no. 3, pp. 747–766, 1972.

[3] J. L. Monteith, "Climate and the efficiency of crop production in britain," *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, vol. 281, no. 980, pp. 277–294, 1977.

[4] M. Marandi, B. Parida, and S. Ghosh, "Retrieving vegetation biophysical parameters and gpp [gross primary production] using satellite-driven lue [light use efficiency] model in a national park," *Environment, Development and Sustainability*, 2022.

[5] C. Beer, M. Reichstein, E. Tomelleri, *et al.*, "Terrestrial gross carbon dioxide uptake: Global distribution and covariation with climate," *Science*, vol. 329, no. 5993, pp. 834–838, 2010.

[6] D. P. Sarkar, B. U. Shankar, and B. R. Parida, "Machine learning approach to predict terrestrial gross primary productivity using topographical and remote sensing data," *Ecological Informatics*, vol. 70, p. 101 697, 2022.

[7] [Online]. Available: `https://fluxnet.org/data/fluxnet2015-dataset/`.

[8] [Online]. Available: `https://modis.gsfc.nasa.gov/data/`.

[9] K. Chandrasekar, M. Sesha Sai, P. Roy, and R. Dwevedi, "Land surface water index (lswi) response to rainfall and ndvi using the modis vegetation index product," *International Journal of Remote Sensing*, vol. 31, no. 15, pp. 3987–4005, 2010.

[10] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, pp. 199–222, 2004.

[11] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[12] J. A. Ohlson and S. Kim, "Linear valuation without ols: The theil-sen estimation approach," *Review of Accounting Studies*, vol. 20, pp. 395–435, 2015.

[13] D. Birkes and Y. Dodge, *Alternative methods of regression*. John Wiley & Sons, 2011.

[14] E. Oikarinen, H. Tiittanen, A. Henelius, and K. Puolamäki, "Detecting virtual concept drift of regressors without ground truth values," *Data Mining and Knowledge Discovery*, vol. 35, no. 3, pp. 726–747, 2021.

[15] [Online]. Available: `https://nannyml.readthedocs.io/en/stable/how_it_works/performance_estimation.html#direct-loss-estimation-dle`.