Improved terrestrial GPP estimation using Multisource Data

Hemant Banke (MD2107)

Under the guidance of Dr. B. Uma Shankar

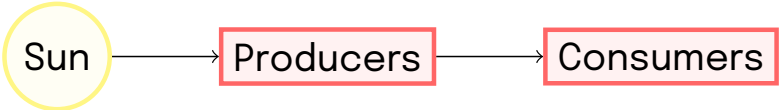
Statistics and Mathematics unit
Indian Statistical Institute, Kolkata
West Bengal - 700 108, India

March 22, 2025

Overview

1. Introduction to GPP
2. Estimating GPP for Australian Region
3. Estimating GPP for Indian Terrain
4. Detecting Concept Drift
5. Direct Loss Estimation
6. Future Work

Gross Primary Productivity (GPP)



- Primary Productivity is the rate at which energy is added to the bodies of primary producers like plants, moss, bacteria, algae, etc.
- GPP $(gC\, m^{-2}\, yr^{-1})$ 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. Producers such as plants use some of this energy for metabolism/cellular respiration and some for growth.
- Net primary productivity (NPP), is GPP minus the rate of energy loss to metabolism and maintenance. It's the rate at which energy is stored as biomass by plants or other primary producers and made available to the consumers in the ecosystem.

Gross Primary Productivity (GPP)

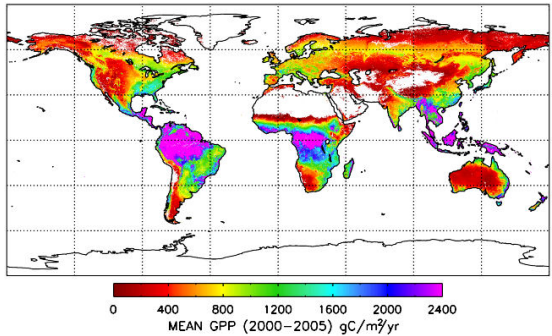


Figure: Mean GPP (between 2000 to 2005) as estimated by MODIS GPP/NPP Project (MOD17) (Source)

Two significant global datasets of terrestrial carbon flux are :

- FLUXNET dataset (measured through Flux Towers)
- MODIS GPP and NPP product, or MOD17 dataset (Remote Sensing dataset)

Estimating GPP 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.

Data for Australian Region

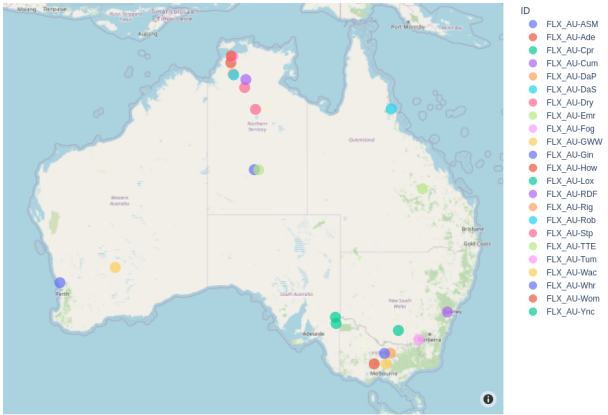


Figure: Map of the 23 Flux tower sites in Australia

For each tower location, the following features were observed between 2001-14.

Features considered for estimation

Biophysical Features (MODIS Remote Sensing dataset)

- *Enhanced Vegetation Index (EVI)* : quantifies vegetation greenness, also corrects for some atmospheric conditions and canopy background noise
- *Leaf Area Index (LAI)* : defined as the one-sided green leaf area per unit ground surface area in broadleaf canopies
- *Fraction of Photosynthetically Active Radiation (FPAR)* : fraction of photosynthetically active radiation (400–700 nm) absorbed by vegetation canopy
- *Land Surface Water Index (LSWI)* : (Computed from Near-Infrared and Short-Wave Infrared bands) sensitive to the total amount of liquid water in vegetation and its soil background.

The spatial resolution is 500m and temporal resolution is 8 days.

Features considered for estimation

Meteorological Features (BOM, Govt. of Australia)

Maximum temperature (Tmax), Minimum temperature (Tmin), Solar Radiation (RAD) and Vapour Pressure (VHP9, VHP15; measured at 9:00 and 15:00 respectively).

The spatial resolution of the gridded product is 0.01 degree and the temporal resolution is of 1 day.

Topographical Features

Elevation and Latitude

To ensure homogeneity in spatiotemporal resolution, the daily products are transformed into their 8-day average.

Data for Australian Region

$\mathbf{X}_{4889\times11}$: At each tower location, the mentioned feature values observed between 2001-14 (measured with 8 day temporal resolution).

$\mathbf{y}_{4889\times1}$: For each tower, the GPP values obtained from *FluxNet2015* dataset, i.e. GPP data measured at Flux Towers is taken as the ground truth. (after 8 day averages to maintain 8 day temporal resolution)

	Latitude	EVI	LSWI	FPAR	LAI	Tmax	Tmin	VHP_9	VHP_15	RAD	ELEVATION	GPP
0	-13.0769	0.362324	0.553240	0.42	1.5	34.562500	24.181250	32.500000	30.166250	21.941250	76	1.531274
1	-13.0769	0.517753	0.529390	0.38	1.0	32.856251	24.639999	33.013748	31.348751	17.456251	76	1.349699
2	-13.0769	0.475762	0.528136	0.57	2.0	32.603752	24.847500	33.562500	31.647501	15.173750	76	1.371809
3	-13.0769	0.361891	0.409836	0.34	1.1	34.107498	24.487499	33.083748	30.233749	19.261250	76	1.410571
4	-13.0769	0.413093	0.476211	0.59	1.9	33.583752	23.646250	32.808750	32.008751	20.917500	76	1.464449

Regression Models Results

Training and Tuning the regression models on $(\mathbf{X}, \mathbf{y})$, using 5-fold cross validation gives following metrics :

Model	RMSE	MAE	R2
Multiple Linear Regression	1.8623	1.2896	71.46%
Support Vector Regression (kernel = 'rbf')	2.6020	2.0248	.
Support Vector Regression (kernel = 'poly', degree = 3, coef0 = 0)	2.8758	2.2722	.
Regression Tree (*max_depth* = 5)	2.1373	1.6026	.
Random Forest (*max_features* = 5, *max_depth* = 8)	1.2660	0.8679	85.70%

Polynomial Regression

For $i \in \{1, 2, ..., 11\}$, we consider $\mathbf{x_i}, \mathbf{x_i^2}, \mathbf{x_i^3}$ as features. Let $\mathbf{Z}_{4889 \times 33}$ be the new data matrix. Then training Multiple regression model on $(\mathbf{Z}, \mathbf{y})$, gives **R-squared = 74.6%, Adj. R-squared = 74.4%**.

Model	R2	Adj. R2
Polynomial Regression	74.6%	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, no multicollinearity are not satisfied. Adding polynomial terms improved the performance but did not help in making the model available for inference, hence we moved away from the linear models setup and tried some non linear models.

Model Evaluation

We choose Random Forest as our final model as it has the best performance metrics.

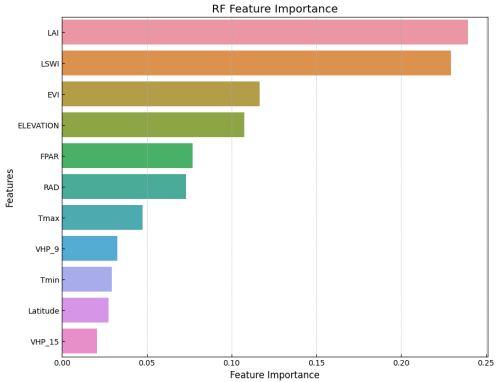


Figure: Feature Importance for RF model

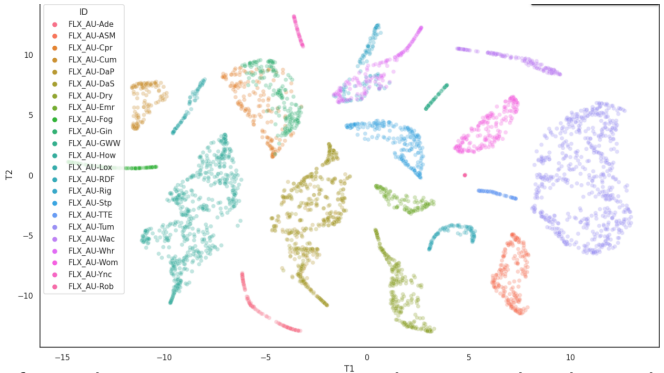
Using data from 2001-13 as Training data, and 2014 as Test data, **Test R-squared is 83.06%, RMSE = 1.3062, MAE = 0.8955**.
Checking MDI based features importance's for this model gives *LAI* and *LSWI* as the most important features followed by *EVI*, *Elevation*, etc. LAI quantifies leaf material in a canopy which directly influences the amount of photosynthesis. LSWI quantifies increase in soil and vegetation liquid water content, which will influence the growth and health of the trees.

Model Evaluation

	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

The RMSE values leaving one tower out suggests that the model performs uniformly across most towers but some towers are worse to predict than the rest. To better inspect this phenomenon we will use t-SNE to visualize our high-dimensional data in a 2-dimensional space.

Model Evaluation



The observations from the same tower are similar to each other. While, observations from different towers are highly dissimilar, barring a few towers. The more dissimilar the observations of a tower are from the rest, the higher is the corresponding LOTO RMSE, as no other tower's observations can explain them effectively.

Estimating GPP for 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.

Modelling Approach

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

Since we don't have enough data to train a good model which works all over India, we will instead use a Global dataset for training the model, so that we have enough data for training.

Modelling Approach

This raises following important questions :

- **(Question 1)** As our training data doesn't contain any observations from Indian region, how do we know if our model can indeed be relied to perform well here ?
- **(Question 2)** How do we estimate the performance of this model ? (since true GPP isn't available for every location in India)

We will deal with these questions later...

Global Dataset

$\mathbf{X}_{5029\times 20}$: At each tower location, the feature values observed between 2001-14 (monthly data).

$\mathbf{y}_{5029\times 1}$: For each tower, the GPP values obtained from *FluxNet2015* dataset. (monthly data)

	LAT	EVI	MIR	NDVI	NIR	BLUE	RED	EASTNESS	NORTHNESS	ELEVATION	...	AET	DEF	PPT	SOIL	SRAD	TMAX	TMIN	VPD	WS	GPP_NT_DT_MEAN
0	47.116669	0.0324	0.0043	0.0755	0.1118	0.1489	0.0961	-0.0293	0.0805	1050.77	...	0.0360	0.0000	11.3052	21.9501	66.4925	1.8304	-6.8503	0.1255	2.4097	0.1044
1	47.116669	0.0283	0.0112	0.1428	0.0492	0.0616	0.0369	-0.0293	0.0805	1050.77	...	6.4343	0.0000	42.6281	21.9585	80.7741	4.4266	-3.7044	0.1718	3.0259	0.2491
2	47.116669	0.3654	0.1431	0.5742	0.2688	0.0285	0.0727	-0.0293	0.0805	1050.77	...	35.1519	0.0000	116.9785	21.9585	149.8898	6.6280	-2.5460	0.2836	2.7027	1.2987
3	47.116669	0.5036	0.0647	0.7637	0.3344	0.0221	0.0448	-0.0293	0.0805	1050.77	...	53.0781	0.1152	58.9471	21.7640	185.3309	7.3712	-1.1729	0.2849	2.5243	6.6094
4	47.116669	0.5690	0.0664	0.8141	0.3689	0.0188	0.0378	-0.0293	0.0805	1050.77	...	82.5099	0.0000	110.4514	21.9585	207.9205	13.8498	4.3432	0.3792	2.5197	13.3325

Fitting Random Forest Model

We consider the forest with 1000 trees and tune the model to find optimal value of hyper-parameter *max_features* i.e. number of features considered to find best split and *max_depth*.

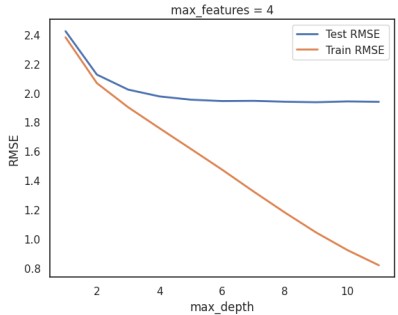


Figure: Training and Test RMSE for *max_features* = 4 and varying *max_depth*

Using 5 fold cross-validation, we obtain Train and Test RMSE (and MAE) values for a grid of hyper-parameters. To prevent overfitting we choose a depth for which the training RMSE is not considerably lower than the test RMSE.

Fitting Random Forest Model

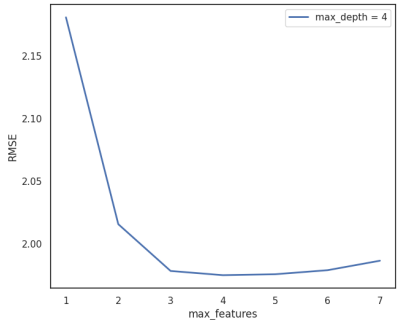


Figure: Test RMSE for *max_depth* = 4 and varying *max_features*

The lowest RMSE is achieved when *max_features* = 4. The same set of optimal values was found through MAE as well.

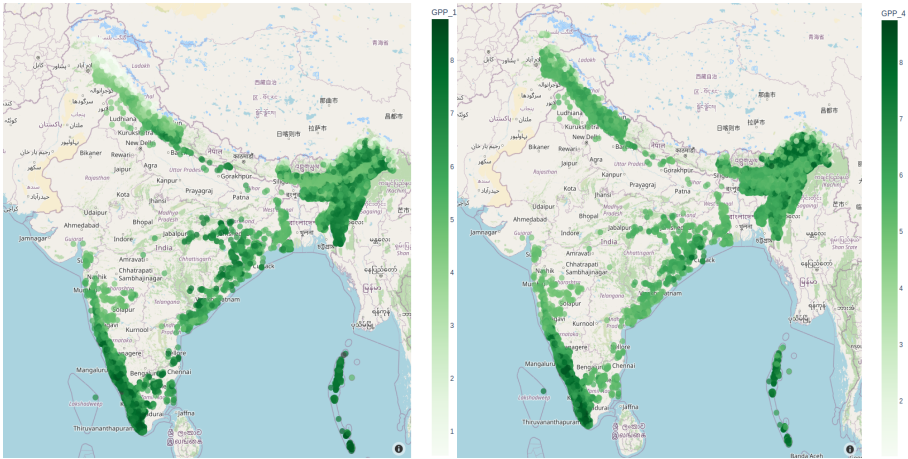
Training Performance :

RMSE	MAE	R-squared
1.7960	1.2092	0.7503

Prediction on Indian Region

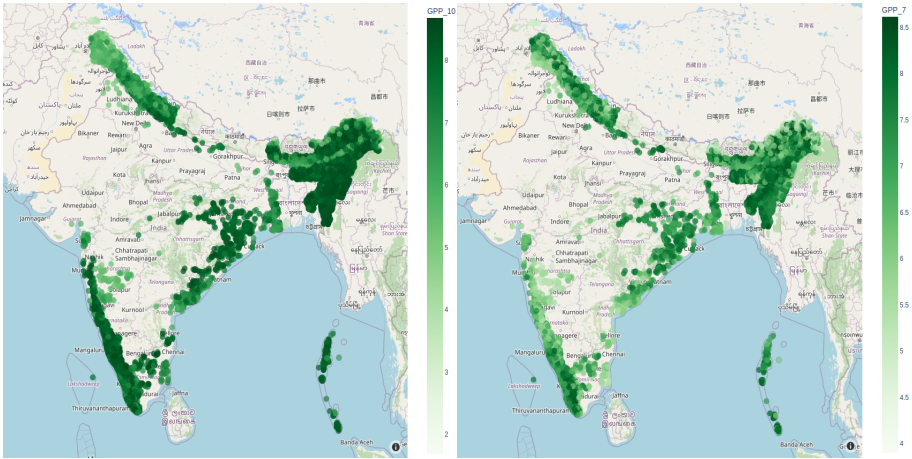
We considered the monthly satellite data of the features over the evergreen forests of India between years 2001-2020. The locations corresponding to Evergreen forests were identified using Land Use/Land Cover (LULC) maps. The satellite data is spatially dense, having resolution of $1km^2$, and contains 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 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.

Prediction on Indian Region (2014)



(a) Winter (b) Spring

Prediction on Indian Region (2014)



(c) Summer (d) Autumn

Prediction on Indian Region - Seasonality (2011-20)

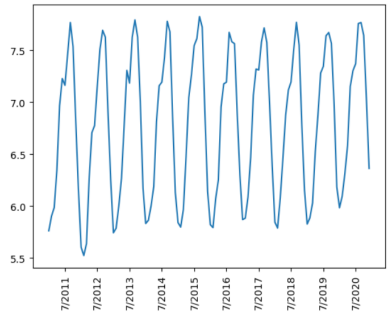


Figure: Trend plot of predicted GPP (2011-20), averaged over all available evergreen locations

The GPP is highest in July(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. Also, the peaks of predicted GPP are slowly shifting forwards, which is inline with the late monsoons due to climate change.

Prediction on Indian Region - Rate of change(2011-20)

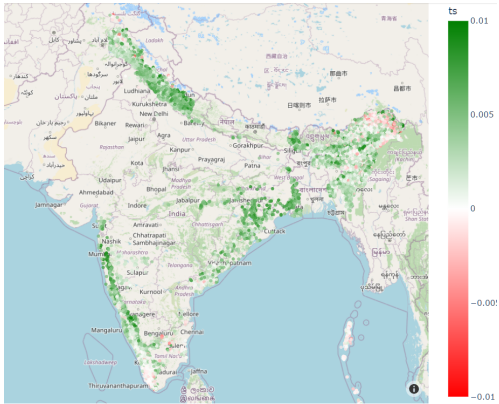


Figure: Rate of change in Predicted GPP from 2011-20, by Theil-Sen estimator

We use the Theil-Sen estimator to find slope instead of the OLS estimate of slope at every location. 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. ($\tau_{TS,OLS} = 0.4785$)

(Question 1) As our training data doesn't contain any observations from Indian region, how do we know if our model can indeed be relied to perform well here ?

Concept Drift

- 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.

Detecting Concept Drift [14]

- Predict if generalization error E of regression 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 and $D_{te} = \{(i, x_i', y_i')\}_{i=1}^{n_{te}}$ be the Test data. Then for a regression function $f : R^m \rightarrow R$ 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

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

Detecting Concept Drift [14]

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 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.

Detecting Concept Drift [14] - Proof

Consider 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'.

Assume we have centered covariates and 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.
We know OLS estimate of β, $\hat{\beta} \sim N(\beta, n^{-1}\Sigma)$.

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

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

Detecting Concept Drift [14] - Proof

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.

Detecting Concept Drift [14] - Training

- 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".
- Train *l* segment models, i.e. model f_i is trained on $D_{|s_i}$, $i = 1(1)l$. The segment models can 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.

Detecting Concept Drift [14] - Testing

- 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 $\mathbf{d = 0.70828}$.

Detecting Concept Drift [14] - Threshold δ

Choosing σ :

- Let f be trained on all segments
- For $i \in \{1, 2, ..., 59\}$, let 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 for 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 $\sigma = \mathbf{2.082}$.

Detecting Concept Drift [14] - Threshold δ

Choosing δ :
Let for $\delta > 0$, we have two classes

$\{\textit{True} = \text{Concept drift Absent in segment}, \textit{False} = \text{Concept drift Present in segment}\}$

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.

Detecting Concept Drift [14] - Threshold δ

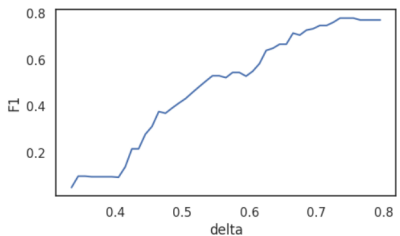


Figure: F1 score versus δ

So we find δ that maximizes the $F1$ score of classifier since a high F1 score indicates that the model has achieved a good balance between precision and recall.

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

The maximum F1 score is achieved at $\delta = \mathbf{0.7551}$.

Hence, 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.

(Question 2) How do we estimate the performance of this model ? (since true GPP isn't available for every location in India)

Direct Loss Estimation [15]

To use the DLE algorithm, we first split our global training data set into train and reference sets of equal sizes using stratified sampling to maintain tower homogeneity.
In DLE algorithm, two models are trained : *child* and *Nanny* model. Child model learns data, while Nanny model learns the loss. Let,

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

Direct Loss Estimation [15] - Training

- 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 $|Y_{ref,j} - \hat{Y}_{ref,j}|$

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

Using RF for both child and nanny models : On the reference set, the child model gives RMSE = 1.7939. The Nanny model has training **RMSE = 0.4348, MAE = 0.2396, R squared = 0.8903**.

Direct Loss Estimation [15] - Testing

- Predict the Child model on Testing set

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

- To get Loss estimate, predict Nanny model as

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

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

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

Comparing this to SVM Model, we get $RMSE = 2.85$ and $MAE = 2.64$.

Future Work

- 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.
- Expanding work to other plant types.

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/`.

References

[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.

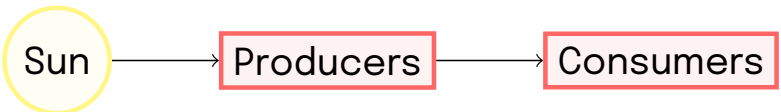
[15] [Online]. Available: `https://nannyml.readthedocs.io/en/stable/how_it_works/performance_estimation.html#direct-loss-estimation-dle`.

The End

Mid Semester Slides

Gross Primary Productivity (GPP)

- GPP $(gC\ m^{-2}\ yr^{-1})$ 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. Producers such as plants use some of this energy for metabolism/cellular respiration and some for growth.
- Net primary productivity (NPP), is GPP minus the rate of energy loss to metabolism and maintenance. It's the rate at which energy is stored as biomass by plants or other primary producers and made available to the consumers in the ecosystem.



(Primary Productivity is the rate at which energy is added to the bodies of primary producers.)

Gross Primary Productivity (GPP)

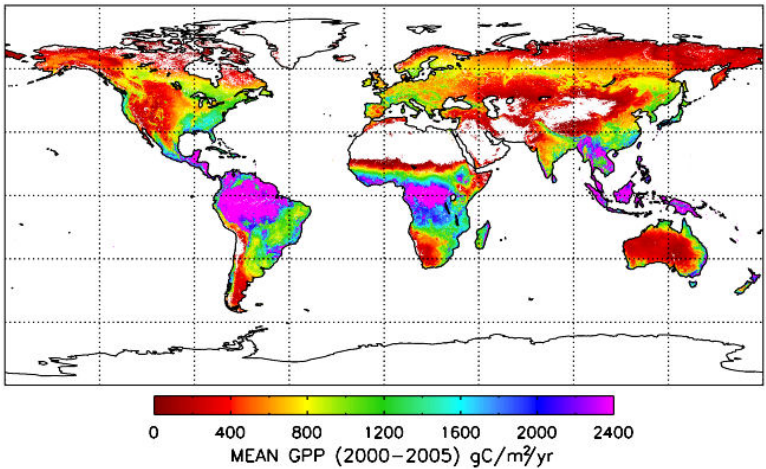


Figure: Mean GPP (between 2000 to 2005) as estimated by MODIS GPP/NPP Project (MOD17) (Source)

Measuring GPP

GPP reflects amount of energy available to support the growth and survival of other organisms within that ecosystem. But measuring GPP is a complex process.

Flux Towers

Flux towers use eddy covariance approach to quantify carbon flux exchange between ecosystem and atmosphere in terms of Net Ecosystem Exchange (NEE: CO_2 fluxes into or out of ecosystem). NEE is then divided into the GPP and ecosystem respiration (RE).

Upscaling data from isolated flux towers is possible using a number of data-driven techniques. Two significant global datasets of terrestrial carbon flux are :

- FLUXNET dataset
- MODIS GPP and NPP product, or MOD17 dataset

Objective

Objective

To predict GPP through machine learning models using Remote Sensing data in combination with meteorological and topographical data for the Australian Region.

Data for Australian Region

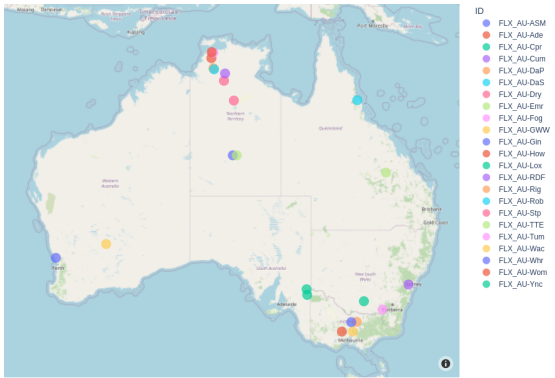


Figure: Map of the 23 Flux tower sites in Australia

GPP data measured at Flux Towers is taken as the ground truth.

Features

The types of Features considered are as follows:

- Biophysical Features (Obtained from MODIS Remote Sensing dataset)
- Meteorological Features (Obtained from BOM, Govt. of Australia)
- Topographical Features

These datasets were downloaded from the year 2001–2014.

Biophysical Features

- *Enhanced Vegetation Index (EVI)* : quantifies vegetation greenness, also corrects for some atmospheric conditions and canopy background noise
- *Leaf Area Index (LAI)* : defined as the one-sided green leaf area per unit ground surface area in broadleaf canopies
- *Fraction of Photosynthetically Active Radiation (FPAR)* : fraction of photosynthetically active radiation (400–700 nm) absorbed by vegetation canopy
- *Land Surface Water Index (LSWI)* : (Computed from Near-Infrared and Short-Wave Infrared bands) sensitive to the total amount of liquid water in vegetation and its soil background.

The spatial resolution is 500m and temporal resolution is 8 days.

Meteorological & Topographical Features

Meteorological Features

Maximum temperature (Tmax), Minimum temperature (Tmin), Solar Radiation (RAD) and Vapour Pressure (VHP9, VHP15; measured at 9:00 and 15:00 respectively).

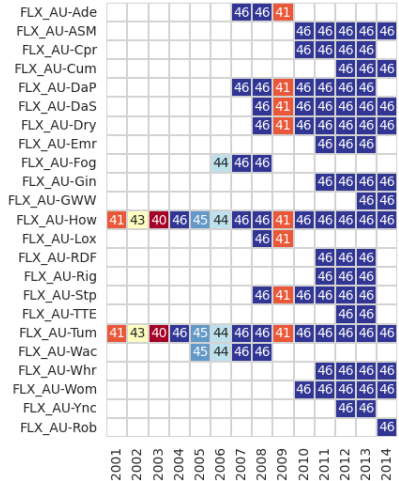
The spatial resolution of the gridded product is 0.01 degree and the temporal resolution is of 1 day.

Topographical Features

Elevation and Latitude

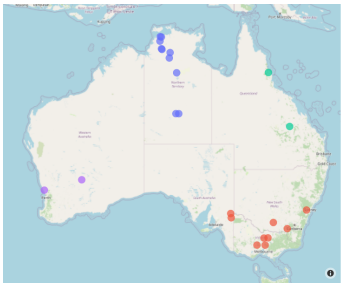
To ensure homogeneity in spatiotemporal resolution, the daily products are transformed into their 8-day average.

Exploratory Analysis

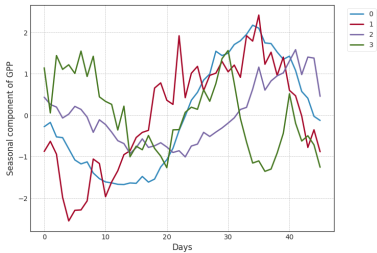


The data contains 4889 observations. Each Tower has atmost 46 observations in a year. But some years do not have all 46 observations. Hence, we can not treat this as a Time-Series problem. Instead we will treat each observation independent.

Seasonal Trend



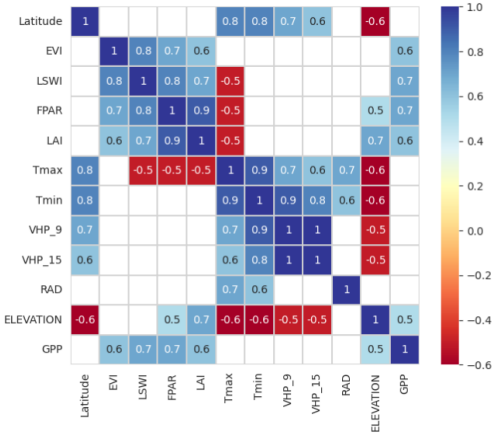
(a) Towers divided into 4 clusters



(b) Seasonal Component of GPP for the 4 clusters

GPP decreases in the beginning of the year, then in-creases and peaks around October after which it starts decreasing again. 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.

Modelling

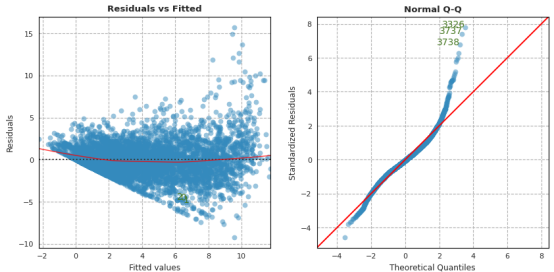


We consider GPP obtained from Flux Towers as the ground truth. This is predicted using 11 features stated before. The Pearson correlation matrix shows presence of linear correlation between GPP and EVI, LSWI, FPAR and LAI. We also notice presence of strong multi-collinearity within features. We will consider the data for years 2001 to 2013 as training set and the year 2014 as our testing dataset.

Fitting Multivariate Linear Regression Model

Training R-squared : 67%

But residuals do not follow Normal distribution, there is presence of heteroscedasticity and strong multicollinearity in our model. So the model defies all assumptions of Linear Regression essential for inference.



Fitting Multivariate Linear Regression Model

To create a better predictive 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.

Improved Training R-squared : 70.1%

Support Vector Regression

Fitting SVR using 10-fold cross validation repeated 3 times, the Mean R-squared and RMSE (with standard deviation) for **Linear, Polynomial** *(degree = 3)*, **Radial basis function (RBF)** are :

	R2	RMSE
rbf	0.758 (0.023)	0.491 (0.037)
poly	0.735 (0.025)	0.513 (0.034)
linear	0.656 (0.024)	0.585 (0.039)

RBF kernel gives us the best results.
Training R-squared : 76.34%

Regression Tree

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 R-squared and Mean RMSE with their standard deviation are **70.8% (0.028)** and **1.890 (0.118)** respectively.
Training R-squared : 73.97%

Random Forest Regression

We consider the forest with 1500 trees and tune the model to find optimal value of hyper-parameter *max_features* i.e. number of features considered to find best split. Using 10 fold cross-validation repeated 3 times, RMSE values for different *max_features* (*max_depth* = 8):

max_features	3	4	5	6	7	8
RMSE	1.729 (0.12)	1.727 (0.12)	1.725 (0.13)	1.726 (0.12)	1.727 (0.13)	1.729 (0.12)

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

The optimal value with Mean RMSE lowest is max features = 5.

Random Forest Regression

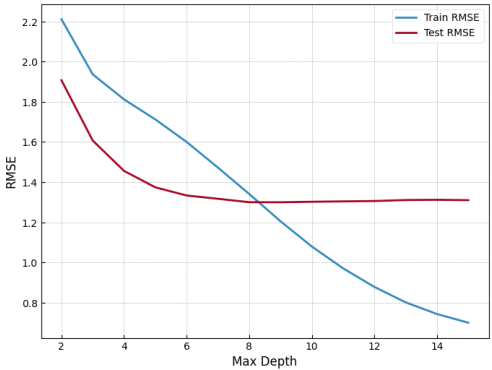


Figure: Training and Test RMSE is plotted for the RF model with *max_features* = 5 and different values of *max_depth*

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.
Training R-squared : 85.70%.

Model Evaluation

We choose Random Forest as our final model as it gives us the highest training R-squared among the rest.
Test R-squared : 83.06%.

Comparing with MODIS GPP estimates

The R-squared between MODIS GPP and FLUXNET GPP over the years 2002 to 2014 is 44.39%, while for year 2014 is 36.41%. So our model outperforms the MODIS GPP estimates and can also upscale GPP from towers to regional scope.

Feature Importance

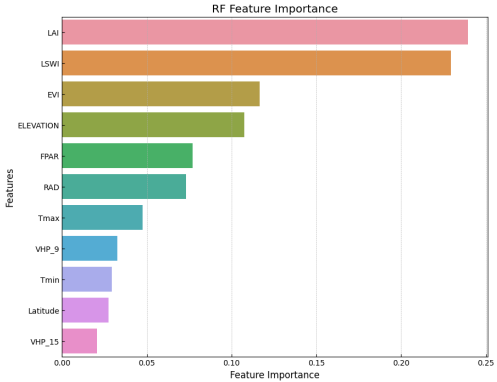


Figure: Feature Importance for RF model

LAI and *LSWI* are the most important features followed by *EVI*, *Elevation*, etc. LAI quantifies leaf material in a canopy which directly influences the amount of photosynthesis possible by the canopy. Land Surface Water Index quantifies increase in soil and vegetation liquid water content, which will influence the growth and health of the trees in the region. So it is justified that these are the most important features found by the model.

LOYO, LOTO

	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

To judge the performance of model over the years and across all Flux Towers, we find the RMSE Leaving One Year Out (LOYO) and Leaving One Tower Out (LOTO). 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 and upgrade of sensors in the Flux Towers.

LOYO, LOTO

	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

The RMSE values leaving one tower out suggests that the model performs uniformly across most towers.

Future Work

- Expanding the work to India which has only 4 Flux towers situated in Haldwani (Uttarakhand), Barkot (Uttarakhand), Meerut (Uttar Pradesh) and Betul (Madhya Pradesh). This involves studying if and how we can expand the model trained using Australian Data on Indian terrain and improve it further using the Fluxnet GPP data from these 4 towers.
- In the absence of ground truth, describing how good our model estimates the true GPP.