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

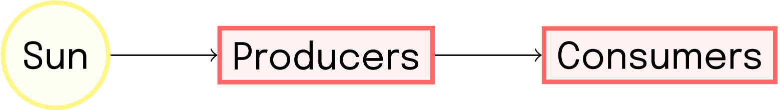
March 22, 2025

Overview

1. **Introduction to GPP**
2. **Objective**
3. **Data Description**
4. **Modelling**
5. **Model Evaluation**
6. **Future Work**

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)

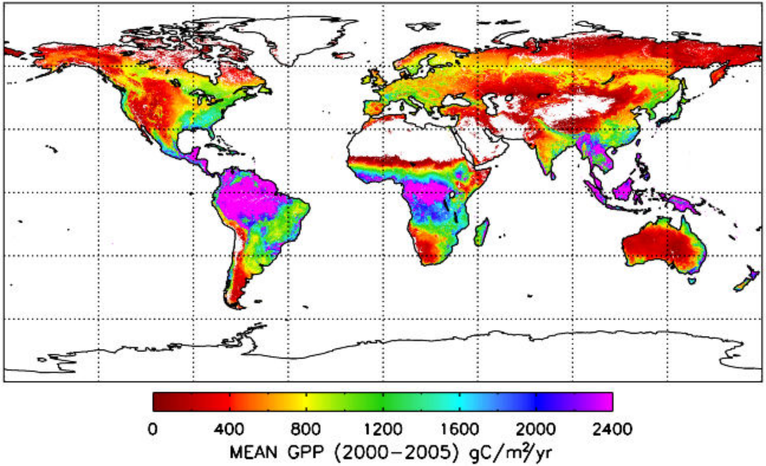


Figure: Mean GPP (between 2000 to 2005) as estimated by MODIS GPP/NPP Project (MOD17) (Source)

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

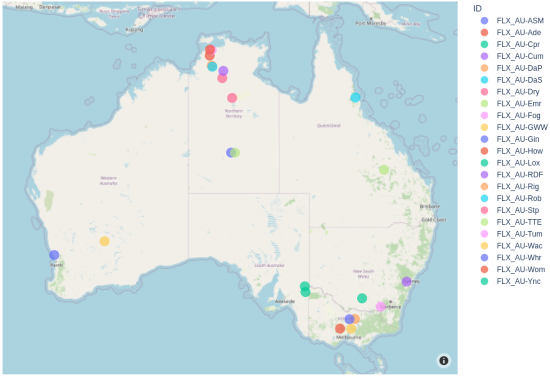


Figure: Map of the 23 Flux tower sites in Australia

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).

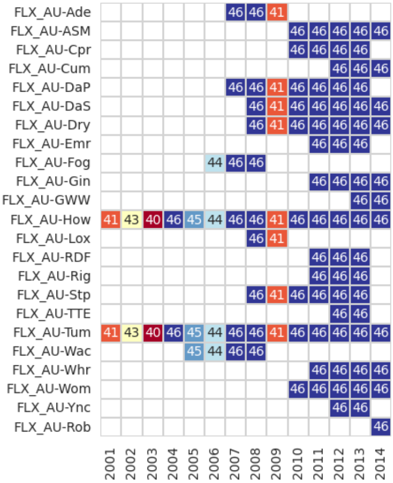
The spatial resolution of the gridded product is 0.01 degree and the temporal resolution is of 1 day.

Topographical Features

Elevation and Latitude

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

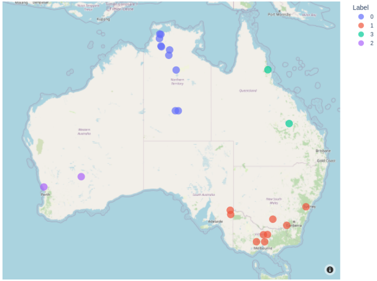
Exploratory Analysis



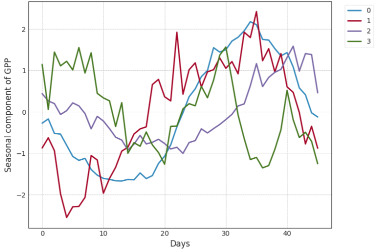
	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014
FLX_AU-Ade							46	46	41					
FLX_AU-ASM										46	46	46	46	46
FLX_AU-Cpr										46	46	46	46	
FLX_AU-Cum												46	46	46
FLX_AU-DaP							46	46	41	46	46	46	46	
FLX_AU-DaS								46	41	46	46	46	46	46
FLX_AU-Dry								46	41	46	46	46	46	46
FLX_AU-Emr											46	46	46	
FLX_AU-Fog						44	46	46						
FLX_AU-Gin											46	46	46	46
FLX_AU-GWW													46	46
FLX_AU-How	41	43	40	46	45	44	46	46	41	46	46	46	46	46
FLX_AU-Lox								46	41					
FLX_AU-RDF											46	46	46	
FLX_AU-Rig											46	46	46	
FLX_AU-Stp								46	41	46	46	46	46	
FLX_AU-TTE												46	46	
FLX_AU-Tum	41	43	40	46	45	44	46	46	41	46	46	46	46	46
FLX_AU-Wac					45	44	46	46						
FLX_AU-Whr											46	46	46	46
FLX_AU-Wom										46	46	46	46	46
FLX_AU-Ync												46	46	
FLX_AU-Rob														46

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



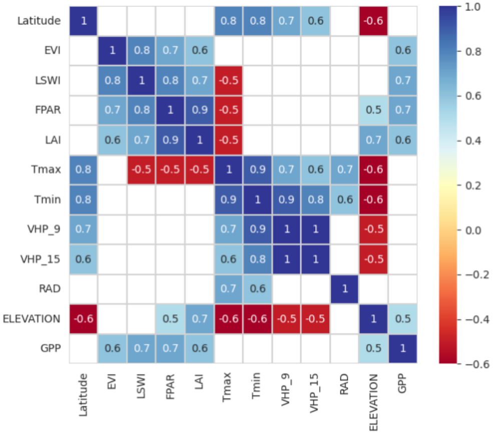
(a) Towers divided into 4 clusters



(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



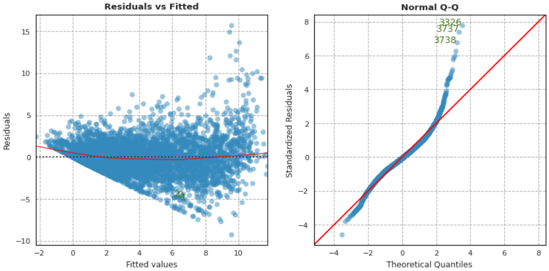
	Latitude	EVI	LSWI	FPAR	LAI	Tmax	Tmin	VHP_9	VHP_15	RAD	ELEVATION	GPP
Latitude	1					0.8	0.8	0.7	0.6		-0.6	
EVI		1	0.8	0.7	0.6							0.6
LSWI		0.8	1	0.8	0.7	-0.5						0.7
FPAR		0.7	0.8	1	0.9	-0.5					0.5	0.7
LAI		0.6	0.7	0.9	1	-0.5					0.7	0.6
Tmax	0.8		-0.5	-0.5	-0.5	1	0.9	0.7	0.6	0.7	-0.6	
Tmin	0.8					0.9	1	0.9	0.8	0.6	-0.6	
VHP_9	0.7					0.7	0.9	1	1		-0.5	
VHP_15	0.6					0.6	0.8	1	1		-0.5	
RAD						0.7	0.6			1		
ELEVATION	-0.6			0.5	0.7	-0.6	-0.6	-0.5	-0.5		1	0.5
GPP		0.6	0.7	0.7	0.6						0.5	1

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.



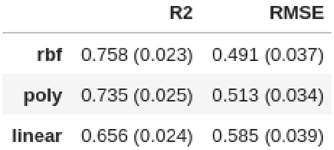
Fitting Multivariate Linear Regression Model

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

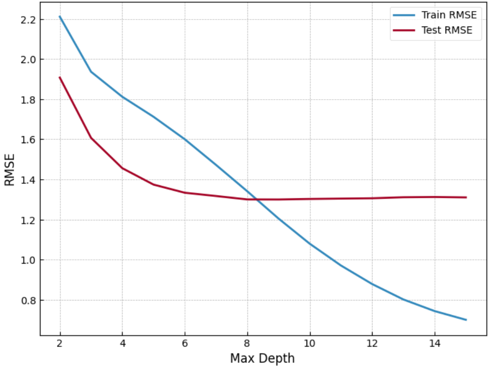


Figure: Training and Test RMSE is plotted for the RF model with *max_features* = 5 and different values of *max_depth*

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

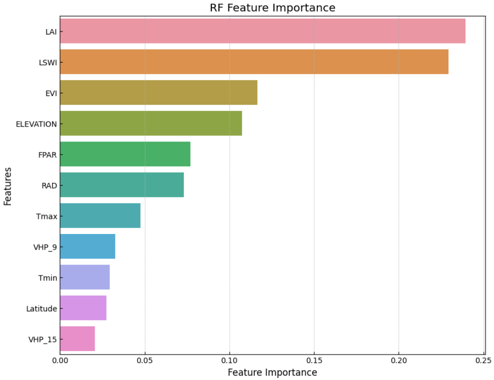
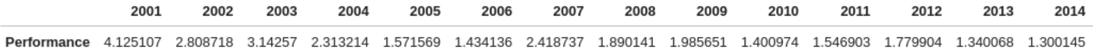


Figure: Feature Importance for RF model

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.

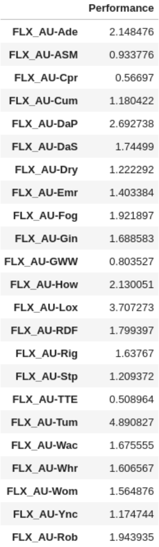
LOYO, LOTO



	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.

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.

The End