Knowledge-Guided Recurrent Neural Networks for Monthly Forest Carbon Uptake Estimation

Andy J.Y. Wong[1]**, Zhicheng Yang**[1]**, Hang Zhou**[1]**, Yuhan Huang**[2]**, Mei Han**[1]**, Jui-Hsin Lai**[1]

[1]PAII Inc., Palo Alto, CA, USA
[2]University of California, Davis, CA, USA
andy.wong.paii@gmail.com, Juihsin.lai@gmail.com

Abstract

To estimate a forest's capability of capturing and storing carbon from the air, modelers generally start with estimating gross primary productivity (GPP). While numerous GPP models have been proposed, this study is among the first in using knowledge of plant growth processes to guide machine learning algorithms for estimating forest GPP with satellite data. When trained and evaluated with flux tower measurements from 47 sites, this approach can accurately estimate monthly GPP with an R^2 of 0.81 and RMSE of 1.52 gC/(m^2d). To better evaluate the proposed model, a comparative experiment against a recent study and an ablation experiment have been conducted. The findings of this study provide a foundation for more robust global monitoring of plant growth and carbon uptakes.

1 Introduction

Covering 30% of Earth's total land areas, forests account for 80% of Earth's total plant biomass and 75% of terrestrial gross primary production (GPP), which is the amount of carbon assimilated by vegetation through photosynthesis (Pan et al. 2013). From 2009 to 2018, forests removed about 30% equivalent of global annual fossil fuel emissions (Friedlingstein et al. 2019). As forest provides a multitude of ecosystem services to our society, many countries are working toward preventing forest degradation, reversing forest cover loss through sustainable forest management, and developing forest carbon offset projects (of Economic and Affairs 2021). Two keys to the success of these efforts are 1. *Quantifying impacts* (e.g. increasing biomass and the amount of carbon absorption) and 2. *Ensuring the impacts are long-lasting*. In the case of forest carbon offset, international standards (Verra 2018) have been developed to quantify carbon absorption. However, the verification methods are labor intensive (Clean Development Mechanism 2011), and most carbon projects lack robust plans to mitigate risks of natural disturbance and climate change (Galik and Jackson 2009; Anderegg et al. 2020; Badgley et al. 2022).

Satellite remote sensing holds promises to reduce the costs of verifying and monitoring forest carbon offsets projects at scale. Due to its ability to continuously and recurrently capture images of the entire earth's surface across the visible, infrared, and microwave spectrum, remote sensing can measure or estimate various physical properties of forests over time (Lechner, Foody, and Boyd 2020), including forest carbon uptakes and stocks. The NASA MOD17A2H(GF) product (Running and Zhao 2021), for example, is widely used in scientific studies on the global and regional carbon flux of GPP (Xiao et al. 2019). The product employs a process-driven approach, specifically the light use efficiency (LUE) model, to estimate GPP. However, its algorithm has a large room for improvement: According to a recent study (Huang et al. 2021), the data product is found to contain significant uncertainties, partly due to uncertainties in model parameters. That study proposed a Markov Chain Monte Carlo approach to calibrate the parameters with measurements from 58 sites, and the improved 8-day estimates explained between 64 to 75% of the variance in measurements (R^2) across four forest types. A related problem with LUE models is that the relationship between leaf biomass (which is quantified with a proxy variable named Leaf Area Index (LAI)) and GPP is one-directional. However, in reality, a surplus or shortage of GPP could in turn induce leaf growth and drop. When the model parameters are uncertain, assuming the LAI-GPP as one-directional could generate incoherent results between GPP of the current time step and LAI of the subsequent time step.

Machine learning, by employing a data-driven approach to estimate GPP using satellite and field measurement data, has the potential to overcome biases presented in process-driven models. Dozens of models have been proposed in the past decade (Joiner and Yoshida 2020; Wei et al. 2017; Tramontana et al. 2016; Stocker et al. 2018), but their interpretability remains a challenge. Inspired by recent research in knowledge-guided machine learning and deep learning (Willard et al. 2021; Liu et al. 2022; Karpatne, Kannan, and Kumar 2022), we encode a part of the Physiological Processes Predicting Growth model (3-PG) (Landsberg and Waring 1997), a process-driven model that has been widely used in forest management (Gupta and Sharma 2019), on PyTorch machine learning framework, where back-propagation is used to update model parameters that have physical meanings. Our contributions are as follows. (1) Our approach of integrating back-propagation into a scientific process-driven model, is the first of its kind in modeling carbon uptake. It could be further expanded to make

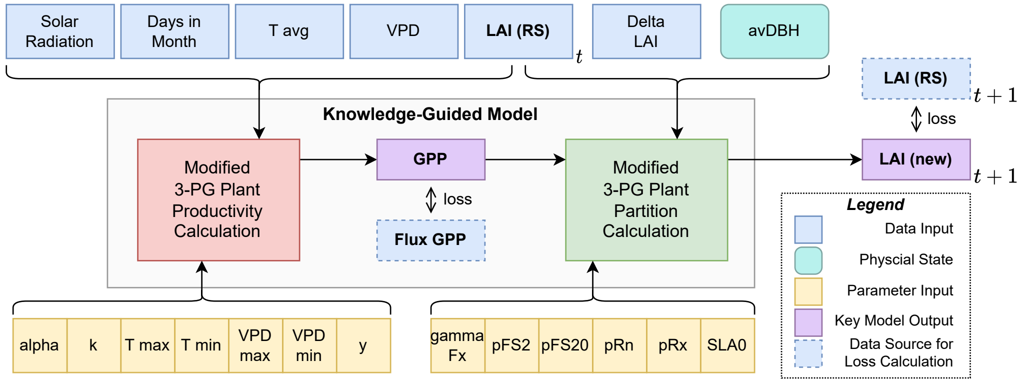

Figure 1: Framework of our proposed model.

use of different emerging satellite data products (e.g. tree counts, density, and height), to estimate carbon stock at different spatial and temporal resolutions, and to accelerate experimentation of alternative models of plant growth processes. (2) Our estimates outperform the calibrated GPP product (Huang et al. 2021) on metrics of R^2 and RMSE in evergreen needle-leaf and deciduous broad-leaf forests, which cover approximately 74% of forests in the United States (Friedl and Sulla-Menashe 2019). (3) Compare to the LUE model used by NASA MOD17A2H(GF) product, 3-PG is more theoretically robust. However due to its complexity, the parameters are traditionally tuned manually. Our research progress demonstrates the feasibility of automated calibration using satellite and field observations, potentially enabling 3-PG to be implemented globally at scale.

2 Materials and Methods

2.1 Dataset

A dataset of pre-processed field measurements and satellite data, covering four forest types: evergreen needle-leaf forest (ENF), evergreen broad-leaf forest (EBF), deciduous broad-leaf forest (DBF), and mixed forest (MF) around the world, has been prepared. The estimated forest age at each site is extracted from the 1km global forest age dataset circa 2010 (Besnard et al. 2021). Monthly field measurements of incoming Shortwave radiation (SW_IN_F [W/m2], converted to solar rad. [MJ/m2/d]), vapor pressure deficit (VPD_F [hPa], converted to VPD [mbar]), and air temperature (TA_F [°C], renamed as T_{avg}) are obtained from the FLUXNET2015 Dataset (Pastorello et al. 2020). The 4-day Leaf Area Index (LAI) (MCD15A3H v6.1) satellite data product (Myneni, Knyazikhin, and Park 2021) are point-sampled with the NASA AppEEARS service (Team 2022). The data are further pre-processed according to Appendix A.1, resulting in 8,648 site-months of data across 98 sites. All aforementioned data have a CC-BY-4.0 license.

2.2 Method

Guided by 3-PG's equations and designs on plant productivity and foliage carbon allocation processes (Landsberg and Waring 1997), we encoded the GPP model in an RNN structure and use back-propagation to tune physically meaningful parameters. The 3-PG's equations are greatly simplified by replacing the complex environmental stress modifiers and removing the age dependency in 3-PG's parameters, but we kept the relationships among key variables intact (e.g. the LAI-GPP bi-directional relationship). The back-propagation process is further constrained through multi-tasks learning and parameter regularization.

Knowledge Guided Machine Learning Layers. Unlike conventional neural networks, where each layer computes an output from a set of inputs and two sets of parameters (e.g. weights for linear and activation functions), layers in this model are encoded according to one of 3-PG's equations (Appendix A.2). A majority of these are linear functions, and a few are simplified from the original equations found in 3-PG. While all layers take at least one input, some layers may have no parameters. As guided by 3-PG's design, the layers are connected, to first estimate monthly plant productivity, including GPP and Net Primary Productivity (NPP), then partition productivity to foliage changes, and finally convert the foliage biomass back to LAI (Fig. 1 and Appendix A.3).

In addition to the core model, we experiment with two modules, namely the canopy quantum efficiency (alpha) modifier and the foliage changes correction modules, in attempt to address some of our model's theoretical weak points. The design of the alpha modifier is analogous to that of the MOD17A2HGF product (Running and Zhao 2021). Its purpose is to reduce the rate of conversion from light to chemical energy, when the plant's environmental conditions deviate from the optimal ranges. Meanwhile, we develop the foliage changes correction module in an attempt to address an overlooked process in the original 3-PG: the seasonality of leaf out and fall (Nölte, Yousefpour, and Hanewinkel 2020).

The number of inputs and parameters required by the model varies by the designs of the model layers. The core model takes four inputs and nine parameters. The inputs are average tree trunk diameter at breast height (avDBH), days in the month, solar radiation, and LAI of the month (RS LAI). In this study, avDBH is assumed to be a constant of 4.57cm; The parameters are specific leaf area at

Parameter	Value	Description
SLA_0	5.63 m2/kg	Specific leaf area at age 0
k	0.50	Extinction coefficient for absorption of photosynthetically active radiation by canopy
alpha	0.06	Canopy quantum efficiency
y	0.47	Ratio of NPP to GPP
gammaFx	0.01 /month	Maximum litterfall rate
pFS_2	1.00	Foliage-stem partitioning ratio at diameter of 2 cm
pFS_{20}	0.29	Foliage-stem partitioning ratio at diameter of 20 cm
pRn	0.25	Minimum fraction of NPP to roots
pRx	0.68	Maximum fraction of NPP to roots
VPD_{min}	6.50 mbar	Minimum threshold for VPD modifiers
VPD_{max}	46.0 mbar	Maximum threshold for VPD modifiers
T_{min}	32.0 °C	Minimum threshold for T_{avg} modifiers
T_{max}	2.00 °C	Maximum threshold for T_{avg} modifiers

Table 1: Initial values of 3PG-related parameters.

age 0 (SLA_0), extinction coefficient for absorption of photosynthetically active radiation by canopy (k), alpha, ratio of NPP to GPP (y), maximum litterfall rate (gammaFx), foliage-stem partitioning ratio at diameter of 2 and 20 cm (pFS_2, pFS_{20}), and minimum and maximum fraction of NPP to roots (pRn, pRx). In addition, the alpha modifier module introduces two extra inputs, VPD and T_{avg}, and four new parameters, which are the minimum and maximum threshold for VPD and T_{avg} modifiers. Meanwhile, the foliage change correction module is driven by the differences in LAI between the current and following month (Delta LAI); This module does not require additional parameters.

Model Architecture. An RNN-like architecture is adopted in this study, where a set of layers is recurrently used to estimate monthly GPP. Unlike conventional RNN, the hidden state is replaced with a "physical" state, which is a subset of variables used and updated by the recurrent layers. In the current version, the model's physical state only consists of avDBH, but it can be expanded to include LAI and other variables (e.g. foliage biomass). Because the only state variable, avDBH, is set as constant, it may be misleading to define the current model as an RNN. Nonetheless, the RNN architecture enables the model to conduct backpropagation through time, in particular, the model can learn from errors it made over a set period of time in a single forward/backward pass, instead of learning from errors from one time step at a time.

Training and Inference. The initial values of parameters are defined by Table 1, and are updated through backpropagation. The ADAM optimizer (Kingma and Ba 2014) is selected for training, with a maximum epoch of 20,000 and a learning rate of 0.001. The process is automatically terminated if training loss (average of 1 epoch) declines by less than 0.001 for 10 consecutive iterations. During inference, the model is given the complete time series of inputs to compute all monthly results at once; During training, the time series data is subsetted with a sliding window, letting the model learn from computing 12 months of estimates at a time. All aforementioned hyper-parameters are arbitrarily chosen, except the learning rate, which is selected based on model performance from a preliminary evaluation.

Loss and Regularization. Although the primary objective of the current model is to estimate GPP, a multi-objective loss (GPP and LAI) is implemented. One reason is that LAI is important for future downstream tasks (e.g. biomass estimation, since GPP that are not used for respiration and growing leaves are generally used to grow tree roots and trunk). Another reason is that we hope the model can leverage the bi-directional GPP-LAI relationship to fine-tune model parameters. Loss is computed as the sum of the squared difference between GPP estimates and measurement, and between LAI estimates and remote sensing LAI of the following month. After computing the weighted sum of GPP and LAI loss (both weight = 1), regularization loss, as defined in Eq. 1, is added for each trained parameter.

$$\begin{aligned}\mathcal{L}_i = &\left(P_i < P_i^{\min}\right) * \left(e^{\left|P_i^{\min} - P_i\right|} - 1\right) * \\ &\left(1/\left(\left|P_i^{\min}\right| + 1e^{-4}\right)\right) + \\ &\left(P_i > P_i^{\max}\right) * \left(e^{\left|P_i^{\max} - P_i\right|} - 1\right)\end{aligned} \tag{1}$$

where P_i refers to previously updated value of parameter i; P_i^{min} and P_i^{max} denote the user-defined minimum and maximum parameter threshold.

Initial Values of Physical Meaningful Parameters. The initial values of 3PG-related parameters are copied from the E. globulus values in the 3PGPJS v2.7 model (Sands 2020). As 3PG equations are simplified in this study, some initial parameter values (SLA_0, gammaFx, pFS_{20}, and pRx) are arbitrarily modified. The initial values of VPD modifier thresholds are copied from the ENF values of Table 2.2 Biome-Property-look-Up-Table (BPLUT) in MOD17 User's Guide (Running and Zhao 2021). The temperature thresholds are arbitrarily defined because our model used average temperature instead of minimum temperature (Running and Zhao 2021) to define thresholds.

3 Experimental Results

Two experiments are conducted in this study using 40 CPUs, and the results are reported in Tables 2 and 3, respectively.

Forest Type	Number of Sites Used (in ours / prev. studies)	Length of Site-Months / Site-8-Days	Core		$GPP_{\text{MC-joint}}$	
			R^2↑	RMSE↓	R^2↑	RMSE↓
ENF	28 (31)	672 / 2576	0.73	1.70	0.65	1.91
EBF	5	120 / 460	0.47	**1.54**	0.73	**1.36**
DBF	10 (16)	240 / 1104	**0.76**	2.27	0.73	2.34
MF	4 (6)	96 / 552	**0.76**	1.80	**0.74**	1.71

Table 2: Results ($gC/(m^2d)$) comparing the proposed model with the calibrated LUE model ($GPP_{\text{MC-joint}}$ (Huang et al. 2021)). (ENF: evergreen needleleaf forest; EBF: evergreen broadleaf forest; DBF: deciduous broadleaf forest; MF: mixed forest.)

Models	Overall (R^2↑ / RMSE↓)	ENF	EBF	DBF	MF
Core	**0.81**/1.52	**0.84/1.28**	**0.40/1.56**	0.78/2.12	0.87/1.23
Core+alpha modifier	0.80/1.55	0.82/1.33	0.13/1.89	**0.81/2.00**	**0.89/1.16**
Core+foliage changes correction	**0.81/1.50**	0.83/**1.28**	**0.40**/1.57	0.80/2.03	0.87/1.24
Full	0.79/1.59	0.82/1.34	-0.05/2.07	0.80/2.02	0.88/1.20

Table 3: Results ($gC/(m^2d)$) comparing four recurrent layers design. (ENF: evergreen needleleaf forest; EBF: evergreen broadleaf forest; DBF: deciduous broadleaf forest; MF: mixed forest.)

The first experiment attempts to compare the performance between the proposed core model and the calibrated LUE model from (Huang et al. 2021). Meanwhile, the second experiment evaluates the incremental improvement of implementing the alpha modifier and the foliage changes correction modules. Following the experimental design of (Huang et al. 2021), we use the same 58 forest sites from (Huang et al. 2021) in our experiments, keep the final two years of data from each site for model evaluation, and use the rest for training. After data pre-processing, 7 sites (IT-SR2, US-Me1, US-Wi9, IT-Isp, IT-PT1, JP-MBF, US-Wi3) have insufficient data points (n¡9) for training; And during training, the model fail to converge at 4 sites (US-Syv, BE-Bra, IT-Ro2, US-WCr). Thus, only 47 sites are available for evaluation. To quickly benchmark our model (monthly estimates, trained site-specifically) against (Huang et al. 2021) (8-day estimates, trained jointly by forest types), we implement two post-processing procedures. Firstly, before inference, we average the trained parameters from the core model by forest types; Secondly, We linearly interpolate the inference results to daily time steps and aggregate them to 8-day averages. In the second experiment, the model is evaluated as is without post-processing.

Despite the unrefined post-processing approach, our model out-perform or is on par with (Huang et al. 2021) over ENF, DBF, and MF sites in the first experiment. The result is encouraging, suggesting the model has the potential to improve the 8-day GPP estimates over most forests in the Earth's temperate region. Further investigation is needed to improve our model over EBF. Perhaps the initial value of parameters also needs to be specified by forest type, or perhaps we over-simplified the model, overlooking important processes that control photosynthesis and growth in EBF.

To our surprise, in the second experiment, the core model outperforms the more complex ones in most cases. The two exceptions occur in DBF and MF, but the improvement is less than 0.03 and 0.12 in terms of R^2 and RMSE. The result suggests the additional modules might not be helpful when parameters are trained specifically by the site. However, they may still be helpful in the future when we attempt to generalize the model by training parameters by forest type. In addition, implementing the foliage changes correction module has significantly improved the agreement between LAI (new) and LAI (RS) (Figure 1), with the overall R^2 and RMSE improving by 0.25 from 0.69 and 0.5 from 0.89, respectively, in the core model. This improvement may be crucial for further downstream tasks, such as estimating foliage biomass and partitioning absorbed carbon to truck, root, and foliage growth.

4 Discussion, Limitations, and Future Work

A real-world application of the NASA GPP product is to be used, together with annual NPP product and vegetation indices from other satellites, to compute a monthly forest sequestration index, which empowers a carbon credit insurance program in China. As our model performs well over ENF, DBF, and MF, which constitute 75% of forests in China (Friedl and Sulla-Menashe 2019), we will integrate it into our insurance application. The proposed model is distinct from the conventional machine learning model in that its design and physical meaningful parameters can be interpreted by forestry experts. And like most deep learning models, our model is incredibly flexible, enabling rapid experimentation and hybrid Physics-ML modeling. However, the current model has several limitations. Firstly, it is constrained by the local minima problem of back-propagation, thus the trained parameter value may not necessarily match the expected "natural" value by experts. Secondly, the model has not been tested for further downstream tasks (e.g. biomass); Its potential to complement the labor-intensive process of verifying and monitoring carbon credit projects has yet to be realized. Here we sincerely invite like-minded researchers to co-develop and co-evaluate this model and its applications.

References

Anderegg, W. R.; Trugman, A. T.; Badgley, G.; Anderson, C. M.; Bartuska, A.; Ciais, P.; Cullenward, D.; Field, C. B.; Freeman, J.; Goetz, S. J.; et al. 2020. Climate-driven risks to the climate mitigation potential of forests. *Science*, 368(6497): eaaz7005.

Badgley, G.; Chay, F.; Chegwidden, O. S.; Hamman, J. J.; Freeman, J.; and Cullenward, D. 2022. California's forest carbon offsets buffer pool is severely undercapitalized. *bioRxiv*.

Besnard, S.; Koirala, S.; Santoro, M.; Weber, U.; Nelson, J.; Gütter, J.; et al. 2021. Mapping global forest age from forest inventories, biomass and climate data. *Earth System Science Data*, 13(10): 4881–4896.

Clean Development Mechanism. 2011. Estimation of carbon stocks and change in carbon stocks of trees and shrubs in A/R CDM project activities. https://cdm.unfccc.int/methodologies/ARmethodologies/tools/ar-am-tool-14-v2.1.0.pdf. Accessed: 2022-09-19.

Friedl, M.; and Sulla-Menashe, D. 2019. MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006. *NASA EOSDIS Land Processes DAAC*.

Friedlingstein, P.; Jones, M. W.; O'Sullivan, M.; Andrew, R. M.; Hauck, J.; Peters, G. P.; et al. 2019. Global Carbon Budget 2019. *Earth System Science Data*, 11(4): 1783–1838.

Galik, C. S.; and Jackson, R. B. 2009. Risks to forest carbon offset projects in a changing climate. *Forest Ecology and Management*, 257(11): 2209–2216.

Gupta, R.; and Sharma, L. K. 2019. The process-based forest growth model 3-PG for use in forest management: A review. *Ecological Modelling*, 397: 55–73.

Huang, X.; Xiao, J.; Wang, X.; and Ma, M. 2021. Improving the global MODIS GPP model by optimizing parameters with FLUXNET data. *Agricultural and Forest Meteorology*, 300: 108314.

Joiner, J.; and Yoshida, Y. 2020. Satellite-based reflectances capture large fraction of variability in global gross primary production (GPP) at weekly time scales. *Agricultural and Forest Meteorology*, 291: 108092.

Karpatne, A.; Kannan, R.; and Kumar, V. 2022. *Knowledge Guided Machine Learning: Accelerating Discovery using Scientific Knowledge and Data*. CRC Press.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Landsberg, J.; and Waring, R. 1997. A generalised model of forest productivity using simplified concepts of radiation-use efficiency, carbon balance and partitioning. *Forest Ecology and Management*, 95(3): 209–228.

Lechner, A. M.; Foody, G. M.; and Boyd, D. S. 2020. Applications in Remote Sensing to Forest Ecology and Management. *One Earth*, 2(5): 405–412.

Liu, L.; Xu, S.; Tang, J.; Guan, K.; Griffis, T. J.; Erickson, M. D.; Frie, A. L.; Jia, X.; Kim, T.; Miller, L. T.; et al. 2022. KGML-ag: a modeling framework of knowledge-guided machine learning to simulate agroecosystems: a case study of estimating N 2 O emission using data from mesocosm experiments. *Geoscientific Model Development*, 15(7): 2839–2858.

Myneni, R.; Knyazikhin, Y.; and Park, T. 2021. MODIS-/Terra+Aqua Leaf Area Index/FPAR 4-Day L4 Global 500m SIN Grid V061. *NASA EOSDIS Land Processes DAAC*.

Nölte, A.; Yousefpour, R.; and Hanewinkel, M. 2020. Changes in sessile oak (Quercus petraea) productivity under climate change by improved leaf phenology in the 3-PG model. *Ecological Modelling*, 438: 109285.

of Economic, U. N. D.; and Affairs, S. 2021. *The Global Forest Goals Report 2021*. United Nations.

Pan, Y.; Birdsey, R. A.; Phillips, O. L.; and Jackson, R. B. 2013. The structure, distribution, and biomass of the world's forests. *Annual Review of Ecology, Evolution, and Systematics. 44(1): 593-622.*, 44(1): 593–622.

Pastorello, G.; Trotta, C.; Canfora, E.; Chu, H.; Christianson, D.; Cheah, Y.-W.; et al. 2020. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. *Scientific Data*, 7(1): 225.

Running, S.; and Zhao, M. 2021. MODIS/Terra Gross Primary Productivity Gap-Filled 8-Day L4 Global 500m SIN Grid V061. *NASA EOSDIS Land Processes DAAC*.

Sands, P. 2020. 3PG-for EXCEL.

Stocker, B. D.; Zscheischler, J.; Keenan, T. F.; Prentice, I. C.; Peñuelas, J.; and Seneviratne, S. I. 2018. Quantifying soil moisture impacts on light use efficiency across biomes. *New Phytologist*, 218(4): 1430–1449.

Team, A. 2022. Application for Extracting and Exploring Analysis Ready Samples (AppEEARS). Ver. 3.10. *NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota, USA*.

Tramontana, G.; Jung, M.; Schwalm, C. R.; Ichii, K.; Camps-Valls, G.; Ráduly, B.; Reichstein, M.; Arain, M. A.; Cescatti, A.; Kiely, G.; et al. 2016. Predicting carbon dioxide and energy fluxes across global FLUXNET sites with regression algorithms. *Biogeosciences*, 13(14): 4291–4313.

Verra. 2018. Verified Carbon Standard. https://verra.org/methodologies/. Accessed: 2022-09-19.

Wei, S.; Yi, C.; Fang, W.; and Hendrey, G. 2017. A global study of GPP focusing on light-use efficiency in a random forest regression model. *Ecosphere*, 8(5): e01724.

Willard, J.; Jia, X.; Xu, S.; Steinbach, M.; and Kumar, V. 2021. Integrating scientific knowledge with machine learning for engineering and environmental systems. *ACM Computing Surveys (CSUR)*.

Xiao, J.; Chevallier, F.; Gomez, C.; Guanter, L.; Hicke, J. A.; Huete, A. R.; et al. 2019. Remote sensing of the terrestrial carbon cycle: A review of advances over 50 years. *Remote Sensing of Environment*, 233: 111383.