Understanding Climate-Vegetation Interactions in Global Rainforests Through a GP-Tree Analysis

Anuradha Kodali[1], Marcin Szubert[2], Kamalika Das[1(✉)], Sangram Ganguly[3], and Joshua Bongard[2]

[1] USRA, NASA Ames Research Center, Moffett Field, CA, USA
anu.uconn@gmail.com, kamalika.das@nasa.gov
[2] University of Vermont, Burlington, VT, USA
{marcin.szubert,jbongard}@uvm.edu
[3] BAERI Inc., NASA Ames Research Center, Moffett Field, CA, USA
sangram.ganguly@nasa.gov

Abstract. The tropical rainforests are the largest reserves of terrestrial carbon and therefore, the future of these rainforests is a question that is of immense importance in the geoscience research community. With the recent severe Amazonian droughts in 2005 and 2010 and on-going drought in the Congo region for more than two decades, there is growing concern that these forests could succumb to precipitation reduction, causing extensive carbon release and feedback to the carbon cycle. However, there is no single ecosystem model that quantifies the relationship between vegetation health in these rainforests and climatic factors. Small scale studies have used statistical correlation measure and simple linear regression to model climate-vegetation interactions, but suffer from the lack of comprehensive data representation as well as simplistic assumptions about dependency of the target on the covariates. In this paper we use genetic programming (GP) based symbolic regression for discovering equations that govern the vegetation climate dynamics in the rainforests. Expecting micro-regions within the rainforests to have unique characteristics compared to the overall general characteristics, we use a modified regression-tree based hierarchical partitioning of the space to build individual models for each partition. The discovery of these equations reveal very interesting characteristics about the Amazon and the Congo rainforests. Our method GP-tree shows that the rainforests exhibit tremendous resiliency in the face of extreme climatic events by adapting to changing conditions.

Keywords: Hierarchical modeling · Symbolic regression
Genetic programming · Earth science · Nonlinear models

1   Introduction

Physics based modeling and perturbation theory has long been used to study the eco-climatic interactions by scientists in order to explain observed phenomena. However, these models, derived under various assumptions of equilibrium, are often only suitable for ideal conditions, and fail to explain the complex dynamics of ecosystem responses to varying environmental factors, especially in the context of a progressively warming global climate. Given the vast amounts of data being collected by different ground-based and remote sensing instruments over long periods of time, the Earth Science research community is extremely data rich. As a result, there has been a slow and steady shift towards the use of machine learning for answering many of their science questions. Ensemble approaches for climate modeling, uncertainty analysis for model evaluation, network based analysis for discovery of new climate phenomena are examples [1]. However, most of the analysis approaches used for climate-vegetation dynamics have been restricted to simple statistical correlation analysis or linear regression [17], thereby limiting discoveries to only linear dependencies. In this work, we formulate the problem of understanding vegetation-climate relationship in rainforests as a regression problem where different climate variables and other influencing factors form the set of independent regressors, and data representing vegetation in the rainforests is the target. In the hope of understanding how climate affects vegetation, we discover regression equations that best fit the observed data. We alleviate the limitation of linear models through the use of a genetic programming based symbolic regression [5] which is a data driven white-box model that allows us to learn both the structure and weights of the regression equation, thereby revealing previously unknown nonlinear interactions in the data. We combine symbolic regression with hierarchical modeling using regression trees in order to partition the large space of spatio-temporal interactions for discovering micro regions within the vast rainforest expanses.

The tropical rainforests are the largest reserves of terrestrial carbon sink, predominantly due to the presence of homogeneous, dense, moist forests over extensive regions. The Amazon forests, for example, are a critical component of the global carbon cycle, storing about 100 billion tons of carbon in woody biomass [7], and accounting for about 15% of global net primary production (NPP) and 66% of its inter-annual variability [19]. Together with the Congo basin in Africa and the Indo-Malay rainforests in Southeast Asia, tropical forests store 40–50% of carbon in terrestrial vegetation and annually process approximately six times as much carbon via photosynthesis and respiration as humans emit from fossil fuel use [6]. With the recent severe Amazonian droughts in 2005 and 2010 [13,17] and on-going multi-decadal drought in the Congo region [20], there is growing concern that these forests could succumb to precipitation reduction, causing extensive carbon release and feedback to the carbon cycle [3]. Interestingly, the two largest rainforests display different characteristic drought patterns with Amazonia encountering episodic and abrupt droughts during the dry season (July–September) and Congo experiencing a gradual and persistent

water shortage. Individual studies of these forests or small areas within them fail to identify any unifying theory that holds for these global rainforests.

In this work we learn from the various observations pertaining to these rainforests in the context of a single modeling framework. We develop a regression tree approach called GP-tree where the models at each node of the tree are built using symbolic regression [5]. This framework discovers dynamics that are local to different partitions within the forests and can be used to explain why certain areas of the rainforests have responded very differently to the extreme climate events of the recent times. The discoveries have been validated by domain scientists conversant with the rainforest ecosystem modeling problem. Precipitation and temperature are the two most relevant climatic factors affecting the rainforests. Other relevant physiological factors that have been included based on domain science expertise are elevation and slope which directly affect how rainfall (or lack thereof) can influence vegetation. Given that forest greenness is an established indicator of tree health, we use satellite-based vegetation greenness observations as our target for this ecosystem model. The goal of the GP tree method is to learn the dependency of greenness on the climatic and physiological factors from historical data spanning multiple years of observations. An additional goal is to identify boundaries in this spatial data set where the equations of dependency change.

2   Related Work

Standard methods in ecosystem modeling use pairwise correlation analysis of vegetation with each climate variable [16]. Trend analysis on standard anomalies of different time series is commonly used for understanding long term dependencies. Nemani et al. [10] use trend analysis for understanding limiting environmental factors in different zones of the earth. Ordinary least squares regression has been used to model the relationship between vegetation and multiple climate variables [8]. Geographic Weighted Regression (GWR) has also been traditionally used to allow for local spatial correlations while explaining climate-vegetation interactions [18]. However, GWR suffers from serious scaling issues. Cubist [11] is another popular analysis tool that automatically partitions the data into geographic regions while learning linear models in each partition. However, none of the methods allow discovery of nonlinear relationships, which severely restricts the discovery process. Nature inspired learning techniques such as deep learning, although very powerful in extracting nonlinear relationships, are not particularly useful in this context due to their blackbox nature.

3   Modeling Framework

Genetic programming based symbolic regression (SR) [5] allows for discovery of nonlinear dependencies in the data by allowing to learn the equation structure along with the regression coefficients. Occasionally when the data is diverse, a single nonlinear model does not suffice. Hierarchical partitioning techniques such

as classification and regression trees (CART) [2] and model trees [11] help in the identification of low variance regions in the data for building individual models. In this paper we describe GP-tree that combines these two powerful algorithms in order to build nonlinear regression models at each partition.

3.1 Symbolic Regression

Symbolic regression's (SR's) main defining features are that it is data driven, white box, and nonlinear. Given training and validation data, SR distills equations of arbitrary form and complexity to explain the data. An example equation explaining vegetation-climate interactions for a specific spatio-temporal extent may look like

$$Y = -0.01log(e^{X_8}(0.03e^{4X_6+X_8+2X_9}((X_5 + X_6)^2 - X_2 - X_3)^2 + 0.2e^{X_{10}}))$$

where $X_i, \forall i$ and Y represent the independent climate variables and greenness respectively. Symbolic regression is instantiated using population-based stochastic optimization method, genetic programming (GP), whose underlying search algorithm is biologically-inspired and consists of 3 major operations, namely, mutation, crossover, and selection [5]. Using these operations, the algorithm iteratively searches the space of possible models by probabilistically recombining previous expressions, modifying their components and adding new random terms to the randomly initialized model population. In each iteration the candidate solutions are evaluated and less accurate and less parsimonious models are replaced by randomly-modified copies of more accurate and more parsimonious models. A squared error measure is used to judge the goodness of fit of the various candidate solutions The set of solutions form a Pareto front where error on the validation set and model complexity are two competing parameters.

3.2 Regression Trees

Decision tree is a machine learning technique for recursively partitioning a space of explanatory (independent) variables in order to better describe a discrete target variable. When the target variables are continuous instead of discrete, regression trees are used. In a regression tree each intermediate node splits the data using a greedy search algorithm that minimizes variance at that node and the leaf nodes contain constant values. A special kind of regression tree called model tree contain leaf nodes which have linear models that can predict the value of a previously unknown example. Regression trees are used in place of a global simple linear regression model where the data has many features that interact in complicated nonlinear ways, and the assumption of linearity falls apart on the entire data set, but might hold true in small subsets. There are different variants of the regression tree algorithms. The original model tree approach proposed by Quinlan [11] relies on building a regression tree with the objective of reducing the standard deviation of the target variable at each split whereas CART [2] chooses to minimize the mean squared error (MSE) of the predicted target value

at each node using decision thresholds. The goodness of fit is determined using the squared error on a validation set and overfitting is handled through tree pruning and cross validation.

3.3   GP-Tree

Our approach, GP-tree consists of two steps: induction of a model tree to partition the data into subsets and then learning of governing equations for each partition using symbolic regression. The overall approach for the GP-tree framework is described in Algorithm 1. The details of the framework are described next.

Algorithm 1. Hierarchical regression: GP-tree

Input: $\mathbf{X} \in \mathbb{R}^{n \times D}, \mathbf{y} \in \mathbb{R}^n, max\_depth, gp\_params$
Output: Tree: **T**, Models: $M_i, i \in k$ (no. of partitions)
Step 1: Build tree: Partition data into k groups
          $\mathbf{T}$ = PolynomialRegressionTree($\mathbf{X}$, $\mathbf{y}$, max_depth)
          $[\mathbf{X_1}, ..., \mathbf{X_k}]$ = Partitiondata($\mathbf{X}$, $\mathbf{T}$)
Step 2: Train GP models
          for each data partition $(\mathbf{X_i}, \mathbf{y_i})$ $(i \in k)$ **do**
              M_i = learnGP($\mathbf{X_i}$, $\mathbf{y_i}$, gp_params)
          end for

Our tree induction differs from the model tree approach in that, instead of the target variance, we consider the MSE approach of CART. Since we are interested in nonlinear models, we compute the MSE for each split using a second order polynomial regression. We hypothesize that the standard deviation of the target variable may not be enough to find homogeneous partitions with respect to the models. In each recursive call of the algorithm (see Algorithm 2), we attempt to find the best binary splitting criterion that divides the dataset $\mathbf{X}$ into two subsets that can be accurately explained by second order polynomial models, which is equivalent of running LASSO on the second order feature combinations of the original data set. To this end, for each feature f we consider a fixed number (100) of scalar threshold values (evenly distributed in the feature domain). For every such pair (feature, threshold) we evaluate the quality of the resulting split by running polynomial regression on the two data subsets $S_1 = \{\mathbf{X}|\mathbf{X_f} < t\}$ and $S_2 = \{\mathbf{X}|\mathbf{X_f} \geq t\}$. The best pair is the one that minimizes the sum of mean squared errors in these subsets. Finally, we invoke the algorithm recursively for the resulting partitions until we reach the maximum depth of the tree. The output of the algorithm is a regression tree with $2^{depth-1}$ internal nodes and 2^{depth} leaves which correspond to partitions of the original dataset. Various methods are available for determining the choice of depth for the model tree [12]; here we use model complexity at the leaf nodes. Although the model tree described above could be used as a predictive model by itself, we attempt to further improve its prediction performance by replacing the second order polynomial models in the terminal leaves of the tree with symbolic regression based models. For each partitions we perform an independent GP

run (see Algorithm 3) using a variant of the Age-Fitness Pareto Optimization (AFPO, [14]) algorithm – a multi-objective method that relies on the concept of genotypic age of an individual (model), defined as the number of generations its genetic material has been in the population. The age attribute is intended to protect young individuals before being dominated by older already optimized solutions.

Algorithm 2. Polynomial Regression Tree

1: **Input:** $\mathbf{X} \in \mathbb{R}^{n \times D}, \mathbf{y} \in \mathbb{R}^n, depth$
2: **Output:** Tree: **T**
3: **if** $depth == 0$ **then**
4:     **return** TerminalNode(LASSO($\mathbf{X}, \mathbf{y}$))
5: **else**
6:     feature, threshold $\leftarrow \arg\min_{f,t}(LR_{error}(\mathbf{X}|\mathbf{X_f} < t, \mathbf{y}) + LR_{error}(\mathbf{X}|\mathbf{X_f} \geq t, \mathbf{y}))$
7:     leftSubtree $\leftarrow$ LinearRegressionTree($\mathbf{X}|\mathbf{X_f} < t, \mathbf{y}, depth - 1$)
8:     rightSubtree $\leftarrow$ LinearRegressionTree($\mathbf{X}|\mathbf{X_f} \geq t, \mathbf{y}, depth - 1$)
9:     **return** InternalNode(feature, threshold, leftSubtree, rightSubtree)
10: **end if**

Algorithm 3. Genetic Programming

1: **Input:** $\mathbf{X} \in \mathbb{R}^{n \times D}, \mathbf{y} \in \mathbb{R}^n, gp\_params$
2: **Output:** GP model: **M**
3: Initialize population of n random models
4: **for** number of generations **do**
5:     Select random parents
6:     Recombine and mutate parents to produce n offspring
7:     Add offspring to the population
8:     Calculate $(error, age, size, complexity)$ for each model in the population
9:     **while** $population\ size > n$ **do**
10:         Select k random models from the population
11:         Determine local Pareto front among k selected models
12:         Remove Pareto-dominated models from the population
13:     **end while**
14: **end for**

The algorithm starts with a population of n randomly initialized individuals each of which has age of one which is then incremented by one every generation. In each generation, the algorithm proceeds by selecting random parents from the population and applying crossover and mutation operators (with certain probability) to produce n offsprings. The offspring is added to the population extending its size to $2n$. Then, Pareto tournament selection is iteratively applied by randomly selecting a subset of individuals and removing the dominated ones until the size of the population is reduced back to n. To determine which individuals are dominated, the algorithm identifies the Pareto front using four objectives (all minimized): prediction error, age, size and expressional complexity. We measure the size of an individual (candidate solution) as the number of nodes in its tree representation. It should be noted here that the regression equation is derived as a tree structure and this tree is different than the hierarchical model tree that is being constructed for the data. For assessing the model complexity, we estimate the order of nonlinearity of the model [15].

4 Data and Computation

MODIS (MODerate-resolution Imaging Spectroradiometer[1]) product MYD13Q1 at 250 m-16day spatio-temporal resolution is used to obtain the Normalized Difference Vegetation Index (NDVI), the most commonly used surrogate for vegetation [9]. Land surface temperature (LST) is similarly derived from MODIS product MYD11A1, but at 1 km-1day spatio-temporal resolution. TRMM (Tropical Rainfall Measuring Mission[2]) observations at 25 km-1month spatio-temporal resolution is used for precipitation measurements. GTOPO30[3] is a global digital elevation model (DEM) at 1 km resolution that is used for obtaining elevation data for the rainforests. Slope is derived from elevation using standard differentials [4]. Since broadleaf evergreens constitute the largest vegetation type found in rainforests, we use a MODIS-derived landcover mask MCD12Q1 to retain only the broadleaf evergreen pixels from the MODIS imagery of the rainforests. All data sets (temporal and spatial resolutions) are selected on the basis of data quality and availability.
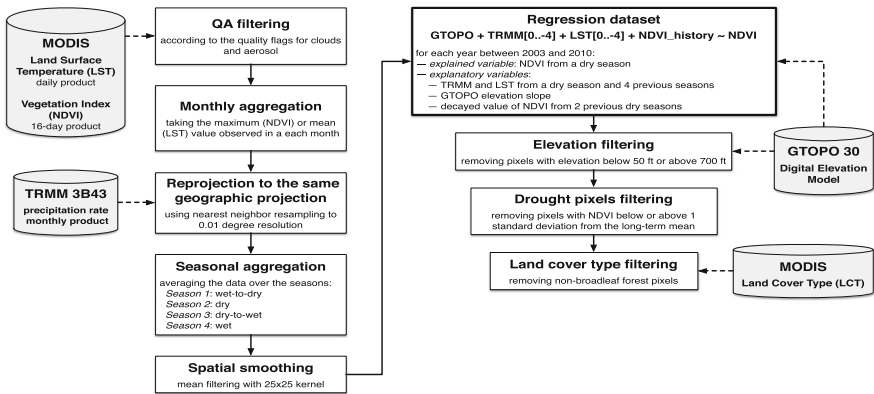

Fig. 1. Data preprocessing pipeline for regression analysis.

For setting up the regression problem, significant amount of preprocessing is needed for colocating and aligning these data products from various sources. Figure 1 shows the end-to-end data preprocessing pipeline. Based on the need of the problem, and the various data sets available, all data sets have been reprojected into the same viewing angle and aligned at 1 km spatial resolution through nearest neighbor interpolation, and averaging based compression. Since seasons largely determine how rainforests respond to environmental influences, we choose a monthly temporal granularity for the study and define the seasons

---

[1] https://modis.gsfc.nasa.gov/.
[2] https://pmm.nasa.gov/trmm.
[3] https://lpdaac.usgs.gov/.

by aggregating monthly time series for each variable as follows: dry season (D) from July to September, dry-to-wet transition (DW) during October, wet season (W) from November to February, and wet-to-dry transition (WD) from March to June. Noise removal is achieved using QA flags available from the MODIS data products. Spatial smoothing over a square neighborhood surrounding each pixel also helps in noise reduction. Land cover filtering indicates removing non-broadleaf pixels while elevation and wetlands filtering removes highly elevated and flooded areas, respectively. Lastly, drought pixels are anomalies with lower vegetation values over years and are removed from the training data.

Regression Setup. Our regression problem is modeling the dry season vegetation as a function of climate and physiological variables in the current (dry) season as well as past seasons going back up to one year. It is set up as follows: $NDVI_k = f(LST_i, TRMM_i, Elev, Slope)$, where $k = current_D$ and $i \in (D_{current}, D_{last}, WD, W, DW)$ are season indices up to one year back in time. The assumption that vegetation in the current season is only affected by rainfall and precipitation within the last one year is based on Subject Matter Expert (SME) feedback and exploratory analysis with different temporal dependencies. We randomly pick 100K examples (out of 700K) from the years 2003–2006 for training our GP-tree model. Year 2007 containing 160K samples is used for validation. The training years chosen using domain knowledge represent drought years and normal years in precipitation. We set the depth of the polynomial decision tree to 2 based on analysis of MSE and model complexity at each leaf node. A tree of depth 2 produces 4 partitions. Once the partitions are obtained using the polynomial regression tree, we spawn the GP optimization routines on each partition with 5000 generations and population size of 50. We use crossover probability of 0.9 and mutation probability of 0.1. Our list of mathematical operations include addition, subtraction, multiplication, logarithm, exponential, square, and cubic. We initialize 30 different optimizations that generate 30 Pareto fronts of GP models. We pick the best model by comparing a subset of models from each front based on size, model complexity, and mean squared error on validation set.

Infrastructure. The data preprocessing pipeline, as well as the modeling and analysis framework have been run on NASA's Pleiades Supercomputer with the following hardware and software configuration. Each of the worker nodes are based on the Intel Sandy Bridge architecture with dual 8 core 2.6 GHz processors and with 32 GB of memory. All nodes' operating systems are running SGI ProPack for Linux kernel version 3.0. Pleiades utilizes a PBS scheduler for job submission. The GP-tree algorithm is centralized and uses a master-slave architecture only for parallelizing the splitting decisions for the various feature-threshold choices (see Sect. 3.3). Once the data is partitioned, the symbolic regression equations are computed at each node using massively parallel search based optimization through genetic programming.

5   Results Analysis

The GP-tree analysis yields 4 different partitions: two of them are temperature limited and precipitation limited zones while two other partitions have a mix of temperature, precipitation, and elevation affecting vegetation. Figure 2 shows the nonlinear equations for each partition. Partitions are identified using blue (leaf 0), cyan (leaf 1), yellow (leaf 2), and red (leaf 3) colors corresponding to the spatial partitions in Fig. 3.

$$leaf0 = 0.43 * LST_{DW} * (-0.13 * Elev^2 + 0.13 * TRMM_{D_{curr}} * LST_{DW}) + 0.06 * TRMM_{D_{last}}^2 - 0.12 * TRMM_{D_{last}}$$
$$* (-LST_{D_{last}} + TRMM_W - (TRMM_{WD} + 0.6 * TRMM_W - 0.26)^2) - 0.09 * LST_{WD}^2 - 0.23 * LST_{WD}$$
$$- 0.23 * LST_W - 0.06 * (Elev + LST_{WD})^2 + 0.16 \tag{1}$$

$$leaf1 = -0.2 * LST_{DW} - 0.4 * LST_{D_{last}} - 0.1 * TRMM_{WD}^2 + 0.1 * TRMM_{WD} + 0.2 * TRMM_W * TRMM_{D_{last}}^2$$
$$* (TRMM_{WD}^2 - TRMM_{WD} - TRMM_W + TRMM_{DW} - LST_{D_{curr}}) + 0.7 * TRMM_W$$
$$- 0.2 * TRMM_{D_{last}}^3 + 0.2 * TRMM_{D_{last}}^2 - 0.3 * LST_{D_{curr}} - 0.3 * LST_{WD} + 0.1 \tag{2}$$

$$leaf2 = 0.05 * TRMM_{WD} - 0.1 * LST_{D_{curr}} - 0.05 * LST_W^2 - 0.2 * LST_W + 0.05 * (-Elev + TRMM_{WD})$$
$$* (-Slope * LST_W + Elev + LST_{WD}) - 0.05 * (-TRMM_{WD} + 0.7)^2 + 0.3 \tag{3}$$

$$leaf3 = -0.12 * LST_W * LST_{DW} - 0.12 * LST_{D_{last}} - 0.02 * TRMM_{D_{curr}}^2 + 0.12 * TRMM_{D_{curr}} + 0.12 * TRMM_{WD}$$
$$- 0.12 * LST_{D_{curr}} - 0.12 * LST_{WD} - 0.12 * LST_W * (TRMM_{D_{last}} + 0.12) - 0.12 * LST_W - 0.12 * log(Elev$$
$$- TRMM_{D_{curr}} + 1.04) - 0.12 * log(log(Elev - 1.09)) \tag{4}$$

Fig. 2. Equations at 4 leaf nodes. Colored boxes indicate matching colors in spatial map in Fig. 3 (Color figure online)
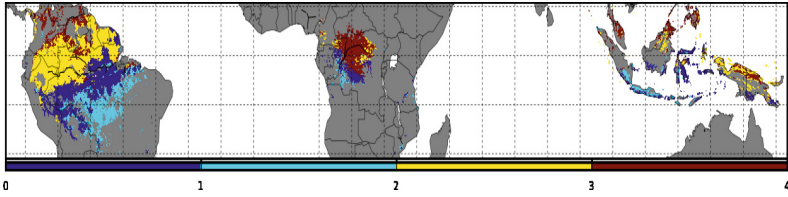

Fig. 3. (a) Partitions of the rainforests obtained through GP-tree (Color figure online)

Figure 3 makes it evident that the Amazonian and African rainforests have characteristically different responses to climate, whereas the Indo-Malay rainforests have no defining nature, comprising of an equal mix of the different partitions. The two main partitions encompassing the bulk of the Amazon river basin are yellow described by Eq. 3 and blue described by Eq. 1 in Fig. 2.

The blue region occupying the central Amazon area is heavily dependent on temperature from the month of October (LST_{DW}), the positive sign indicating that vegetation in that area prefers colder temperatures during the dry to wet season transition. The presence of the $TRMM$ terms in Eq. 1 indicates vegetation dependence on seasonal rainfall as well. It shows resilience since a relatively dry wet season (low rainfall during November–February) is compensated by a wetter transition and vice versa. it also shows that vegetation in this
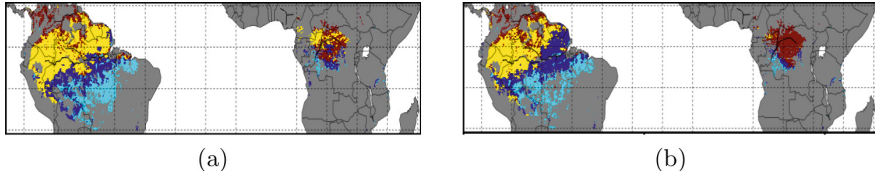
Fig. 4. Partitioning (a) 2005 and (2010) pixels of Amazon and Africa using learned GP-tree model (Color figure online)

region does not thrive in excessive rainfall. This can be explained as an effect of the interruption of the adiabatic cooling process that forces temperatures to rise in extreme cloud conditions, thereby effecting vegetation negatively. The yellow partition in the north of the Amazon governed by Eq. 3 requires colder temperatures along with longer rainfall spells overflowing from the wet season to the transition season for increased greening of the trees. The cyan and red partitions representing Eqs. 2 and 4 respectively are spread across the peripheral regions of the Amazon basin. The southern periphery (cyan region) is heavily dominated by wet season rainfall, as seen in Eq. 2. A similar cyan area can also be seen flanking the southern Congo basin Africa. Geographically, both these regions represent a transitional zone in the rainforests, where there is a mix of broadleaf evergreens and savannas (grasslands) that completely depend on rainfall for greening. On the other hand, it is apparent that bulk of the African forests is governed by Eq. 4 described in red in Fig. 3. This is the most complex model including precipitation and temperature covariates from almost all seasons. Lack of copious rainfall in this region for the last two decades has ruined all seasonal patterns for the broadleaf evergreens as they try to sustain themselves through the low to moderate rainfall received during all seasons, while relying on lower temperatures in this region.

These equations enable domain scientists to explain several observations made in the last decade about these rainforests. Given the dependence of any rainforest on appropriate rainfall and temperatures, the permanent state of drought in the African Congos in the last 15 years have led the trees in that region to gradually succumb to the drought indicated by a decreasing NDVI trend [20] over the years. Even slight improvement in rainfall in certain years results in those trees trying to adapt to a different steady state behavior, evident from the appearance of yellow patches in the African red partition in Fig. 4a. The Amazon droughts of 2005 and 2010 also manifest themselves similarly. The trees in the drought-stricken regions of the Amazon, in an attempt to survive under these extreme climatic conditions, adapt to a different steady state behavior (a different equation). As seen in Fig. 4a, a large part of the blue river basin region affected by the 2005 drought turns yellow to account for the sudden water deficiency through increased photosynthetic activity [13]. Similarly, a small part of the yellow region near the mouth of the Amazon river becomes blue after the 2010 drought hits that area, thereby resisting tree dieback due to the unfavorably low rainfall and high temperatures caused by the El Niño phenomenon in

that year. This study shows how the global rainforests, although suffering from frequent droughts and rising temperatures, generally show very strong resilience by adapting to changing conditions.

Model Performance. We compare performance of the GP-tree model with 4 different baselines: (i) a single linear model, (ii) a single symbolic regression model, (iii) linear regression tree with linear models at the leaves, and (iv) polynomial regression tree with linear models. We compare mean squared error on a standard validation set (examples for year 2007) for each model. The MSEs are shown in Table 1. The progressive improvement of error as we go from linear to nonlinear model, and from a single global model to multiple models obtained through hierarchical partitioning is evident from the error values. Our method improves the state of the art (first baseline) by almost 43%.

Table 1. Table showing mean squared error for GP-tree and the baseline methods for ecosystem modeling

GP-tree	Baseline 1	Baseline 2	Baseline 3	Baseline 4
0.28	0.49	0.31	0.45	0.38

6 Conclusion

For ages, scientists have been trying to understand the effect on climate and other environmental variables on vegetation. Given that the rainforests are the largest carbon sinks, it is particularly important to understand how these forests react under changing climatic conditions, and whether their future is at risk. Existing studies using simple correlation analysis or linear regression models built at a global level, have failed to capture the nuanced dependencies of vegetation in micro regions within these rainforests on environmental factors. In this study we use genetic programming based approach symbolic regression for discovering equations that model the vegetation climate dynamics in the rainforests of the world. Expecting micro-regions within the rainforests to have unique characteristics compared to the overall general characteristics, we hierarchically partition the space using a regression tree approach called GP-tree and nonlinear regression models for each partition. Our GP-tree framework discovers that these rainforests exhibit very different characteristics in different regions. We also see that in the face of extreme climate events, the trees adapt to reach a different steady state and therefore, exhibit resiliency.

References

1. Banerjee, A., Monteleoni, C.: Climate change: challenges for machine learning. Tutorial at NIPS 2014 (2014)
2. Breiman, L., Friedman, J., Stone, C., Olshen, R.: Classification and Regression Trees. Taylor & Francis, Milton Park (1984)
3. Cox, P.M., Betts, R.A., Jones, C.D., Spall, S.A., Totterdell, I.J.: Acceleration of global warming due to carbon-cycle feedbacks in a coupled climate model. Nature **408**(6809), 184–187 (2000)
4. Horn, B.: Hill shading and the reflectance map. IEEE Proc. **69**, 14–47 (1981)
5. Koza, J.R.: Genetic Programming: On the Programming of Computers by Means of Natural Selection, vol. 1. MIT Press, Cambridge (1992)
6. Lewis, S.L., et al.: Increasing carbon storage in intact african tropical forests. Nature **457**(7232), 1003–1006 (2009)
7. Malhi, Y., et al.: The regional variation of aboveground live biomass in old-growth amazonian forests. Glob. Change Biol. **12**(7), 1107–1138 (2006)
8. Mao, K., et al.: Estimating relationships between NDVI and climate change in Quizhou province, Southwest China. In: 2010 18th International Conference on Geoinformatics, pp. 1–5, June 2010
9. Myneni, R., Hall, F., Sellers, P., Marshak, A.: The interpretation of spectral vegetation indexes. IEEE Trans. Geosci. Remote Sens. **33**(2), 481–486 (1995)
10. Nemani, R.R., et al.: Climate-driven increases in global terrestrial net primary production from 1982 to 1999. Science **300**(5625), 1560–1563 (2003)
11. Quinlan, J.R.: Learning with continuous classes. In: Proceedings of the Australasian Joint Conference on Artificial Intelligence, pp. 343–348. World Scientific, Singapore (1992)
12. Rokach, L., Maimon, O.: Top-down induction of decision trees classifiers - a survey. IEEE Trans. Syst. Man. Cybern. Part C (Appl. Rev.) **35**(4), 476–487 (2005)
13. Saleska, S.R., Didan, K., Huete, A.R., Da Rocha, H.R.: Amazon forests green-up during 2005 drought. Science **318**(5850), 612–612 (2007)
14. Schmidt, M., Lipson, H.: Age-fitness Pareto optimization. In: Riolo, R., McConaghy, T., Vladislavleva, E. (eds.) Genetic Programming Theory and Practice VIII. GEVO, vol. 8, pp. 129–146. Springer, New York (2011). https://doi.org/10.1007/978-1-4419-7747-2_8
15. Vladislavleva, E.J., Smits, G.F., den Hertog, D.: Order of nonlinearity as a complexity measure for models generated by symbolic regression via pareto genetic programming. IEEE Trans. Evol. Comput. **13**(2), 333–349 (2009)
16. Xiao, J., Moody, A.: Geographical distribution of global greening trends and their climatic correlates: 1982–1998. Int. J. Rem. Sens. **26**(11), 2371–2390 (2005)
17. Xu, L., Samanta, A., Costa, M.H., Ganguly, S., Nemani, R.R., Myneni, R.B.: Widespread decline in greenness of Amazonian vegetation due to the 2010 drought. Geophys. Res. Lett. **38**(7) (2011)
18. Yuan, F., Roy, S.: Analysis of the relationship between NDVI and climate variables in minnesota using geographically weighted regression and spatial interpolation, vol. 2, pp. 784–789 (2007)
19. Zhao, M., Running, S.W.: Drought-induced reduction in global terrestrial net primary production from 2000 through 2009. Science **329**(5994), 940–943 (2010)
20. Zhou, L., Tian, Y., Myneni, R.B., Ciais, P., Saatchi, S., et al.: Widespread decline of congo rainforest greenness in the past decade. Nature **509**, 86 (2014)