Understanding climate-vegetation interactions in global rainforests through a GP model-tree analysis

Anuradha Kodali[1]*, Marcin Szubert[2], Sangram Ganguly[3],
Joshua Bongard[2], and Kamalika Das[4]

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

Abstract. The tropical rainforests are the largest reserves of terrestrial carbon sink 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 ongoing drought since 2000 in the Congo region there is growing concern that these forests could succumb to precipitation reduction, causing extensive carbon release and feedback to the carbon cycle. Contradicting research has claimed that these forests are resilient to such extreme climatic events. A significant reason behind these diverse conclusions is the lack of a holistic spatio-temporal analysis of the remote sensing data available for these regions. Small scale studies that use statistical correlation measure and simple linear regression to model the climate-vegetation interactions have suffered from the lack of complete data representation and the use of simple (linear) models that fail to represent physical processes accurately, thereby leading to inconclusive or incorrect predictions about the future. In this paper we use a genetic programming (GP) based approach called symbolic regression for discovering nonlinear 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 and build a nonlinear GP model for each partition. The discovery of these equations reveal very interesting characteristics about the Amazon and the Congo rainforests. Overall it shows that the rainforests exhibit tremendous resiliency in the face of severe droughts. Based on the partitioning of the observed data points over years, we can conclude that in the absence of adequate precipitation, the trees adopt to reach a different steady state and recover as soon as precipitation is back to normal.

Keywords: hierarchical modeling, symbolic regression, genetic programming, earth science, nonlinear models

* This work was done when the author was at NASA Ames

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 [18], thereby limiting discoveries to only linear dependencies. In this work, we formulate the problem of understanding vegetation-climate relationship in rainforests as a nonlinear 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 GP based regression method, called symbolic regression. The strength of the approach lies in the fact that it learns both the structure and weights of the regression equation and is, therefore, able to identify previously unknown nonlinear interactions in the data representing physical processes of interest. We combine symbolic regression with hierarchical modeling using decision 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 Amazonian forests, for e.g., 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 interannual variability [20]. Together with the Congo basin rainforests 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 [14,18] and on-going drought since 2000 in the Congo region [21], there is growing concern that these forests could succumb to precipitation reduction, causing extensive carbon release and feedback to the carbon cycle [3]. Contradicting research claims that these forests are resilient to such extreme climatic events [12]. The point of interest is that these 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 since 1985. Even within each rainforest, very different reactions to these droughts are being observed starting from greening [14], to mortality [13], leading to controversies arising from contradicting claims about the future of the rainforests.

In this paper we tie the various observations pertaining to these rainforests into a single modeling framework through the use of a hierarchical regression tree-based approach called GP model-tree where the models at each leaf node of the tree are built using GP based symbolic regression [5]. This framework discovers physical processes that are local to different partitions within the forests and can explain why certain areas of the rainforests have responded very differently to the extreme climate events of the recent times. These processes are represented by complex nonlinear terms identified by the GP search and have been validated by domain scientists conversant with the problem. The problem that we focus on is the dependency of the health of the rainforests on different climatic factors, namely, precipitation and temperature. Since the greenness of trees is an indicator of whether a tree is thriving, we use a satellite based vegetation index as a surrogate for representing vegetation health. We also add other relevant factors such as elevation and slope which directly affect how rainfall (or lack thereof) can influence vegetation in an area. In addition to describing the GP model-tree framework and the discoveries made, this paper also provides insights into the data preprocessing challenges that are unique to this domain and provides a principled approach for preprocessing these multimodal remote sensing data sets.

1.1 Competing methods

Standard methods used in this domain for understanding climate-vegetation dependencies include pairwise correlation analysis of vegetation with each climate variable [17]. Trend analysis by calculating standard anomalies of different time series is the most common practice to understand temporal and spatial variations. Nemani et al. [10] use this analysis for understanding limiting environmental factors in different zones of the earth. While these studies help in understanding global trends and the strength of the linear relationship with each climate variable, regression models predict the relationship of vegetation with more than one variable. Ordinary least squares have been used to model the relationship between vegetation and climate variables [8]. Geographic Weighted Regression (GWR) has also been traditionally used to allow for local spatial effect while explaining climate-vegetation interactions [19]. However, this method suffers from serious scaling issues. Cubist regression is another method that automatically partitions the data into regions while learning linear models in each partition [11]. However, none of these methods allow discovery of complex nonlinear relationships and are therefore not useful in discovery of physical processes in the ecosystem. In the next section we describe our GP model-tree framework. It should be noted here that deep learning based approaches, although very powerful in unveiling nonlinear relationships cannot be used in this context because an important aspect of this equation discovery process is the ability of the domain scientist to understand and explain the physical meaning of the equation, that a blackbox model will not be able to produce.

2 Technical Approach

GP based symbolic regression [5] allows for discovery of unknown physical processes by allowing to learn the equation structure along with regression coefficients. However, a single global model is often not enough in the presence of spatio-temporal variations in the data. Therefore, we use hierarchical partitioning of the data along the lines of classification and regression trees (CART) [2]. Each terminal node in the tree represents a unique nonlinear relationship specific to points in that partition.

2.1 Symbolic regression

Symbolic regression's (SR's) main defining features are that it is data driven, white box, and nonlinear. It is data driven in the sense that the investigator needs to provide only training and validation data; SR will distill equations with arbitrary form and complexity to explain the data. An example equation explaining vegetation climate interactions for a specific spatio-temporal extent can 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 represents the independent environmental variables. Symbolic regression is typically instantiated using a population-based stochastic optimization method called GP as the underlying search algorithm is biologically-inspired. In short, terms are randomly added, removed or modified to individual models, and less accurate and less parsimonious models are replaced by randomly-modified copies of more accurate and more parsimonious models. Such an approach has the major drawback of requiring considerable computational effort since learning a good equation is a stochastic search process that requires generating and testing many thousands (and sometimes even millions or more) of candidate solutions. Some variant of a squared error measure is used to judge the goodness of fit of the various candidate solutions.

2.2 GP model-tree

Our approach, GP model-tree consists of two steps. We first induce a tree to divide the space into partitions and then we learn the governing equations for each partition using symbolic regression. The overall approach for the GP model-tree framework is described in Algorithm 1. The details of the framework are described next.

In the first step we induce a *model tree* – a special case of a regression tree in which each terminal node contains a model that is used to produce the final prediction value. The original model tree approach proposed by Quinlan

[11] relies on building a traditional regression tree with standard deviation used as an impurity measure that allows to determine the best split variable and split threshold [2]. Only after such a tree is built, the constant values in leaves are replaced with linear regression models fitted to the data in each leaf. In our algorithm, we adopt the mean squared error of a second order polynomial regression model as the impurity measure because the data well explained by this model may have high standard deviation.

Algorithm 1 GP model-tree

```
Input: $\mathbf{X} \in \mathbb{R}^{n \times D}, \mathbf{y} \in \mathbb{R}^n, max\_depth, gp\_params$
Output: Tree: $\mathbf{T}$, Models: $\boldsymbol{M_i}, i \in k$ (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
        $\boldsymbol{M_i}$ = learnGP($\mathbf{X_i}$, $\mathbf{y_i}$, $gp\_params$)
    end for
```

The model tree is constructed in a traditional greedy, top-down, divide-and-conquer manner. Note that, we have used polynomial factors of order 2 for this purpose in our analysis. 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. To this end, for each feature f we consider a fixed number of scalar threshold values (evenly distributed in the feature domain). For each such pair (feature, threshold) we evaluate a 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 the 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.

Algorithm 2 Polynomial Regression Tree

```
Input: $\mathbf{X} \in \mathbb{R}^{n \times D}, \mathbf{y} \in \mathbb{R}^n, depth$
Output: Tree: $\mathbf{T}$
if $depth == 0$ then
    return TerminalNode(LASSO($\mathbf{X}, \mathbf{y}$))
else
    feature, threshold ← $\arg\min_{f,t}(LR_{error}(\mathbf{X}|\mathbf{X_f} < t, \mathbf{y}) + LR_{error}(\mathbf{X}|\mathbf{X_f} \geq t, \mathbf{y}))$
    leftSubtree ← PolynomialRegressionTree($\mathbf{X}|\mathbf{X_f} < t, \mathbf{y}, depth - 1$)
    rightSubtree ← PolynomialRegressionTree($\mathbf{X}|\mathbf{X_f} \geq t, \mathbf{y}, depth - 1$)
    return InternalNode(feature, threshold, leftSubtree, rightSubtree)
end if
```

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 leaves of the tree with models produced by GP based symbolic regression. For each of the partitions identified by the model tree, we perform independent randomized GP runs (see Algorithm 3). For this purpose we use a variant of the Age-Fitness Pareto Optimization (AFPO, [15]) algorithm – a multiobjective 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 3 Genetic Programming

```
Input: $\mathbf{X} \in \mathbb{R}^{n \times D}, \mathbf{y} \in \mathbb{R}^n, gp\_params$
Output: GP model: $\mathbf{M}$
Initialize population of $n$ random models
for number of generations do
    Select random parents
    Recombine and mutate parents to produce $n$ offspring
    Add offspring to the population
    Calculate $(error, age, size, complexity)$ for each model in the population
    while $population\ size > n$ do
        Select $k$ random models from the population
        Determine local Pareto front among $k$ selected models
        Remove Pareto-dominated models from the population
    end while
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 decision tree that is being constructed for the data. For assessing the expressional complexity, we estimate the order of nonlinearity of the model [16].

3 Data sets and processing

We use satellite-based measurements for vegetation, precipitation, temperature during years 2000-2010 and digital elevation model (DEM) measurements for elevation to learn steady state equations for the broadleaf evergreen forests in relation to environmental variables. Data products from the twin MODerate-resolution Imaging Spectroradiometer (MODIS) sensors aboard NASA's Earth observation System (EOS)-era Aqua satellite are used for vegetation, temperature, and land cover masks. Normalized Difference Vegetation Index (NDVI) [9], a surrogate for vegetation is obtained from the MODIS product MYD13Q1 that provides 250-meter sinusoidal projected surface reflectance data adjusted using a bidirectional reflectance distribution function (BRDF) and collected at intervals of every 16 days. Similarly, land surface temperature (LST) is obtained from the MYD11A1 product at 1KM spatial resolution collected daily during the day. Broadleaf evergreen forests are identified using masking information available in the MCD12Q1 data product for land cover. The Tropical Rainfall Measuring Mission (TRMM) launched jointly by NASA and Japan Aerospace Exploration Agency provides monthly precipitation measurements derived from the combination of 3B42 products (every 3-hours data product) and Global Precipitation Climatology Centre (GPCC) rain gauge analysis at 25KM spatial resolution. GTOPO30[1] is a global digital elevation model (DEM) with a spatial resolution of 30 arc seconds (approximately 1KM) that provides elevation height and derived slope [4] used in the regression study. All data sets (temporal and spatial resolutions) are selected on the basis of data quality and availability.

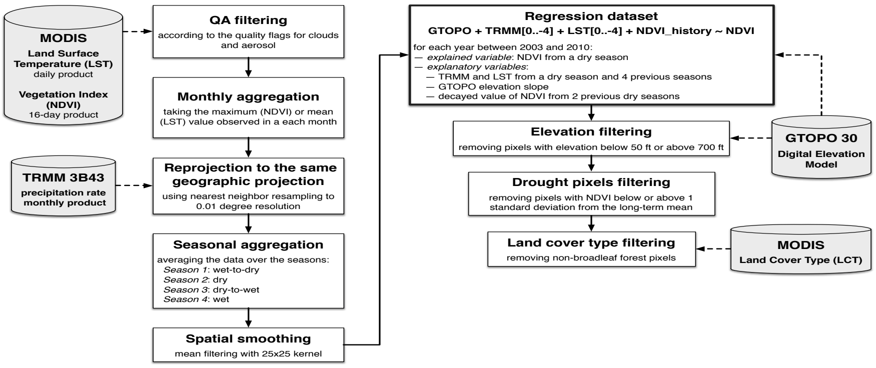

MODIS
Land Surface Temperature (LST)
daily product
Vegetation Index (NDVI)
16-day product
QA filtering
according to the quality flags for clouds and aerosol
Monthly aggregation
taking the maximum (NDVI) or mean (LST) value observed in a each month
TRMM 3B43
precipitation rate monthly product
Reprojection to the same geographic projection
using nearest neighbor resampling to 0.01 degree resolution
Seasonal aggregation
averaging the data over the seasons:
Season 1: wet-to-dry
Season 2: dry
Season 3: dry-to-wet
Season 4: wet
Spatial smoothing
mean filtering with 25x25 kernel
Regression dataset
GTOPO + TRMM[0..-4] + LST[0..-4] + NDVI_history ~ NDVI
for each year between 2003 and 2010:
— explained variable: NDVI from a dry season
— explanatory variables:
— TRMM and LST from a dry season and 4 previous seasons
— GTOPO elevation slope
— decayed value of NDVI from 2 previous dry seasons
Elevation filtering
removing pixels with elevation below 50 ft or above 700 ft
GTOPO 30
Digital Elevation Model
Drought pixels filtering
removing pixels with NDVI below or above 1 standard deviation from the long-term mean
Land cover type filtering
removing non-broadleaf forest pixels
MODIS
Land Cover Type (LCT)

Fig. 1: Data preprocessing pipeline for regression analysis.

For setting up a regression problem, we first need to preprocess these data sets to reproject them in the same viewing angle, and align them with respect to spatial and temporal resolutions. All data sets are reprojected to 1KM (0.01x0.01 degree) resolution while aggregating to monthly-level by averaging measurements in a calendar month. These monthly values are cumulatively averaged to four seasons per year, namely., dry season (D) during July to September, dry-to-wet transition (DW) during October, wet season (W) during November to February, and the wet-to-dry transition (WD) during March to June. First level of noise removal is achieved using QA flag based filtering during retrieval from the MODIS

[1] https://lpdaac.usgs.gov/

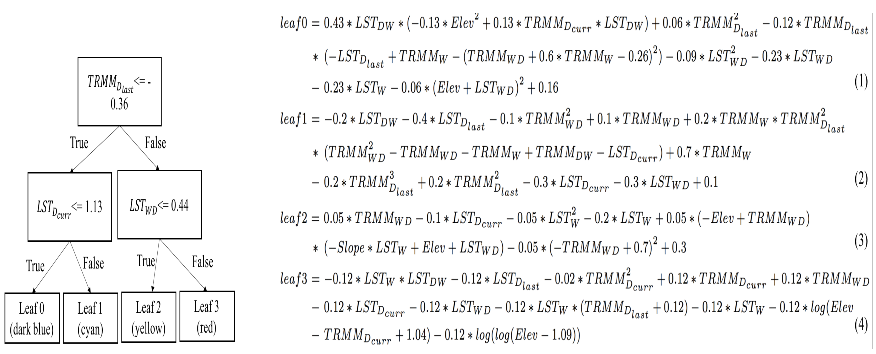
data products. In order to reduce noise further, we perform spatial smoothing by averaging measurements in an adjoining square grid around every pixel. Land cover filtering is done to remove pixels not representing the broadleaf category, while elevation and wetlands filtering removes highly elevated and flooded areas, respectively. Lastly, drought pixels are anomalies with lower vegetation values over years and are removed from the training data. Figure 1 describes the process in details.

Problem setup Once the data is prepared, we set up the regression problem 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 dry 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 settings. From our data set we pick years 2003 to 2007 for building our GP model-tree model with 100K examples randomly chosen every 10 generations to evaluate the training error. 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 500. We use crossover probability of 0.9 and mutation probability of 0.1 [5]. 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. The models that are obtained almost always contain nonlinear terms which is an indication that linear regression models are not enough to capture the complex physical processes in climate-vegetation interactions. The average normalized mean squared error for the GP model-tree is 0.31 and the maximum improvement seen for any year of test data over linear regression based decision tree is more than 16%.

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.

4 Discoveries

The goal of this analysis is discovery of equations that best explain the vegetation observations in the global rainforests, given recorded climate data and other environmental factors such as land elevation, and slope. The GP model-tree analysis yields 4 different partitions that broadly divides the global rainforests into temperature limited, and precipitation limited zones. In addition we find two more zones that have a mix of temperature, precipitation, and elevation affecting vegetation. These are mostly transitional forest regions that border



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

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

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

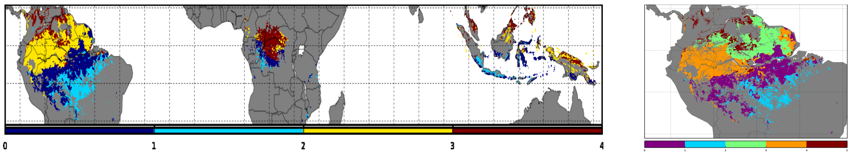
$$\begin{aligned} leaf3 = {} & -0.12 * LST_W * LST_{DW} - 0.12 * LST_{D_{last}} - 0.02 * TRMM^2_{D_{curr}} + 0.12 * TRMM_{D_{curr}} + 0.12 * TRMM_{WD} \\ & - 0.12 * LST_{D_{curr}} - 0.12 * LST_{WD} - 0.12 * LST_W * (TRMM_{D_{last}} + 0.12) - 0.12 * LST_W - 0.12 * log(Elev \\ & - TRMM_{D_{curr}} + 1.04)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

(a) (b)

Fig. 2: (a) Polynomial regression tree, (b) Equations at each leaf node

the main forested areas in the two biggest rainforests in the Amazon and the Congo region. Further partitioning the space yields micro regions within these big partitions that have unique characteristics that explain the behavior of these forests in the face of droughts and recent events of extreme water shortage. In this section we describe in details the main climatic influencers of these different regions.

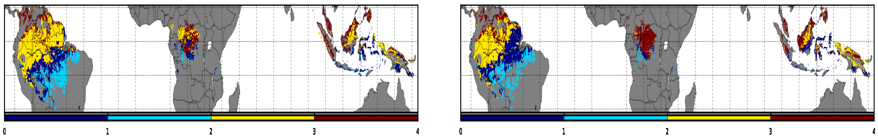
Figure 2a shows the decision tree partitions obtained by running our algorithm on the global rainforest data set and Figure 2b shows the nonlinear equations for each of these partitions. Partitions are identified using blue (leaf 0), cyan (leaf 1), yellow (leaf 2), and red (leaf 3) colors, as shown in Figure 3a and are conditioned upon precipitation during last year's dry season at level one and temperature at level two of the regression tree. Note that in all equations the target is NDVI of the current year's dry season, and being a measure of greenness captured, it is acting as a surrogate for vegetation in the area.



(a) (b)

Fig. 3: (a) Global partitions, (b) Subdivision of leaf 0 (blue) and leaf 2 (yellow) pixels of (a) into 4 partitions. Image best viewed in color.

Looking at the partitions in Figure 3a, it is evident that the Amazonian rainforests and the African rainforests have characteristically different response to climate, whereas the Indo-Malay rainforests have no defining nature, and comprises of an equal mix of the different partitions. The two main partitions encompassing the bulk of the Amazon river basin are yellow described by Equation 3 and blue described by Equation 1 in Figure 2b. They are dependent on both temperature and rainfall across different seasons with temperature being the most dominant feature. The yellow region in northern Amazon requires colder



(a) (b)

Fig. 4: Repartitioning using derived tree for (a) 2005 pixels and (b) 2010 pixels. Image best viewed in color.

temperatures along with longer rainfall spells overflowing from the wet season to the transition season for increased greening of the trees. The blue region occupying the central Amazon river basin also prefers colder temperatures during the wet seasons. However, they also are dependent on cross-seasonal rainfall patterns. A low rainfall wet season is easily compensated for by a wetter transition and vice versa. However, too much rainfall is not good for the trees to flourish in this region. This is due to interruption of the adiabatic cooling process of the region which forces temperatures to rise and negatively affect vegetation. On further partitioning the points in the blue and yellow region in our data sets, we see that these big regions in the Amazon are partitioned broadly into three regions: the purple region in Figure 3b which mostly overlaps with the blue central river basin and the orange and lime regions which encompass most of the yellow partition. These partition equations (not shown here due to lack of space) correspond to the geographic features of the region and include elevation, and dry season rainfall as additional environmental influencers. On the other hand, it is apparent that bulk of the African forests is governed by Equation 4 described by the color red in Figure 3a. The dominant climate variables in this equation are temperature from all seasons and extended wet season precipitation. The biggest reason for such behavior is the lack of copious rainfall in these regions during any time of the year. This leads to the trees trying to sustain themselves through the low to moderate rainfall received during other seasons, and also through lower prevalent temperatures. The cyan region, described by Equation 2, is very heavily controlled by wet season rainfall and flanks the southern border of both the Amazonian and the African rainforests. Geographically, these regions represent a transitional zone in the rainforests, where there is a mix of broadleaf evergreens and the bordering savannas (grasslands). Clearly, this region has a unique nature due to the influence of the savannas, that transcends continental boundaries.

These equations enable scientists to explain several observations made in the last decade about these rainforests. Given the dependence of the African Congos on good rainfall and low 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, as is seen through a decreasing NDVI trend [21] 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 Figure 4a. The Amazon droughts of 2005 and 2010 also sees similar behavior in that 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. As seen in Figure 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 [14]. 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 El Niño in that year. The general increase of cyan pixels in 2010 can be attributed to increased deforestation activities that has been plaguing the southern Amazon for the past decade. This analysis explains to a great extent the contradicting observations and conclusions drawn by various studies that either look at the rainforests at a macroscopic level or analyze small regions that fail to capture the global picture.

5 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 nuances of the micro regions that exist within these rainforests and respond to the climatic changes very differently. In this study we use a GP based approach called 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 polynomial regression-tree based hierarchical partitioning of the space and build a nonlinear GP model for each partition. Our 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 adopt to reach a different steady state and therefore, exhibit resiliency.

Acknowledgments. This research is supported in part by the NASA Advanced Information Systems Technology (AIST) Program's grant (NNX15AH48G) and in part by the NASA contract NNA-16BD14C. The authors would also like to thank Dr. Ramakrishna Nemani, who is an expert in this area, for his insightful comments and perspective on some of the research findings.

References

1. Banerjee, A., Monteleoni, C.: Climate change: Challenges for machine learning. Tutorial at NIPS'14 (2014)

2. Breiman, L., Friedman, J., Stone, C., Olshen, R.: Classification and Regression Trees. Taylor & Francis (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 (Jan 1981)
5. Koza, J.R.: Genetic programming: on the programming of computers by means of natural selection, vol. 1. MIT press (1992)
6. Lewis, S.L., Lopez-Gonzalez, G., Sonké, B., Affum-Baffoe, et al.: Increasing carbon storage in intact african tropical forests. Nature 457(7232), 1003–1006 (2009)
7. Malhi, Y., Wood, D., Baker, T.R., Wright, J., Phillips, O.L., Cochrane, et al.: The regional variation of aboveground live biomass in old-growth amazonian forests. Global Change Biology 12(7), 1107–1138 (2006)
8. Mao, K., Li, M., Chen, C., Huang, Q., Chen, Z., Li, F., Chen, D.: Estimating relationships between ndvi and climate change in guizhou 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. Geosci. and Remote Sensing, IEEE Trans. on 33(2), 481–486 (1995)
10. Nemani, R.R., Keeling, C.D., Hashimoto, H., Jolly, W.M., Piper, S.C., Tucker, C.J., Myneni, R.B., Running, S.W.: 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 Aus. Joint Conf. on Artificial Intelligence. pp. 343–348. World Scientific, Singapore (1992)
12. Sakschewski, B., von Bloh, W., Boit, A., Poorter, L., Peña-Claros, M., Heinke, J., Joshi, J., Thonicke, K.: Resilience of Amazon forests emerges from plant trait diversity. Nature Climate Change 6, 1032–1036 (Nov 2016)
13. Salazar, L.F., Nobre, C.A., Oyama, M.D.: Climate change consequences on the biome distribution in tropical south america. Geophys. Res. Letters 34(9) (2007)
14. 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)
15. Schmidt, M., Lipson, H.: Age-Fitness Pareto Optimization. In: Genetic Programming Theory and Practice VIII, Genetic and Evolutionary Computation, vol. 8, pp. 129–146. Springer New York (2011)
16. 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. on Evol. Comp. 13(2), 333–349 (2009)
17. Xiao, J., Moody, A.: Geographical distribution of global greening trends and their climatic correlates: 1982–1998. Int. J. of Rem. Sens. 26(11), 2371–2390 (2005)
18. 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. Geophysical Research Letters 38(7) (2011)
19. 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)
20. Zhao, M., Running, S.W.: Drought-induced reduction in global terrestrial net primary production from 2000 through 2009. Science 329(5994), 940–943 (2010)
21. 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)