Predicting the future of the Amazon rainforests using regression analysis

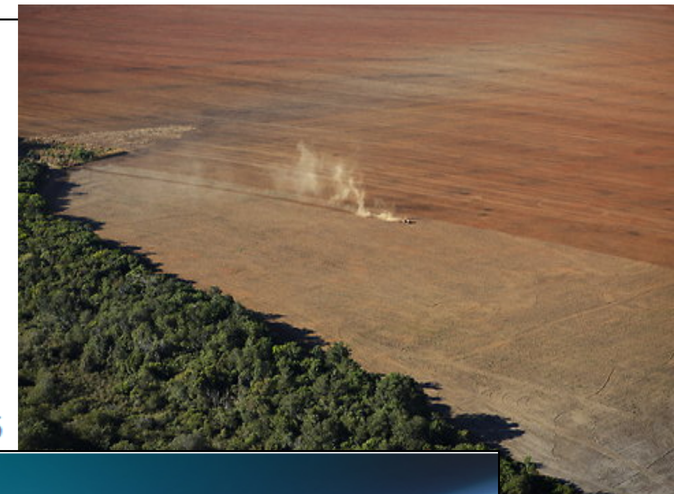
Kamalika Das
NASA Ames Research Center

Joint work with: Anuradha Kodali[+], Sangram Ganguly[+], Marcin Szubert[-], Josh Bongard[-]
+ NASA Ames Research Center
- University of Vermont

Problem Definition

- Understand magnitude and extent of ecosystem exposure, sensitivity and resilience to the 2005 and 2010 Amazon droughts
- Understand the effect of climatic factors on vegetation anomalies using regression analysis

Motivation



nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue

Long-term decline of the Amazon carbon sink

Nature **519**, 344–348 (19 March 2015) | doi:10.1038/nature14283
Received 09 April 2014 | Accepted 04 February 2015 | Published online 18 March 2015

nature climate change

Home | Current issue | News & comment | Research | Archive ▾ | Authors & referees ▾ | About the jo

Projected strengthening of Amazonian dry season by constrained climate model simulations

Nature Climate Change **5**, 656–660 (2015) | doi:10.1038/nclimate2658
Received 06 November 2014 | Accepted 17 April 2015 | Published online 01 June 2015

nature climate change

Home | Current issue | News & comment | Researc

Temperature as a potent driver of regional forest drought stress and tree mortality

Nature Climate Change **3**, 292–297 (2013) | doi:10.1038/nclimate1693
Received 21 March 2012 | Accepted 28 August 2012 | Published online 30 September 2012

Problem formulation

- Point-to-point regression analysis
- Estimate spatio-temporal dependency of forest ecosystems on climate variables

$$V_{ij}^{t}=f(Lc_{ij}, E_{ij}, CV_{ij}^{t}, CV_{nb}^{t}, CV_{ij}^{t-1}, CV_{nb}^{t-1}, \ldots\ldots CV_{ij}^{t-k}, CV_{nb}^{t-k})$$

V:vegetation,
LC:landcover type,
CV:climate variable(s)
E: elevation

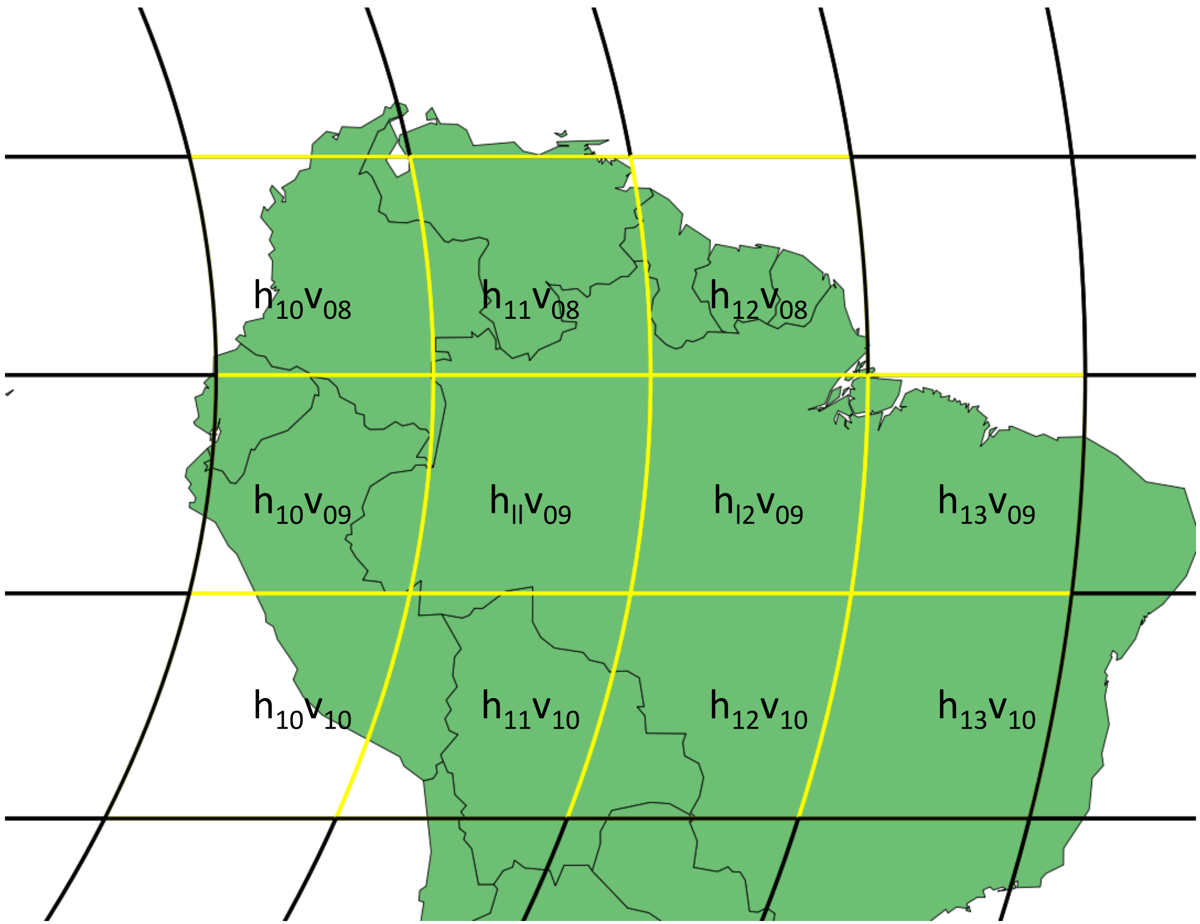
i,j: pixel location indices
t: time index
nb: spatial neighborhood of index i,j
k: temporal dependency

Open challenges: 1. Estimating function *f*
2. Estimating best choices for *k, nb*

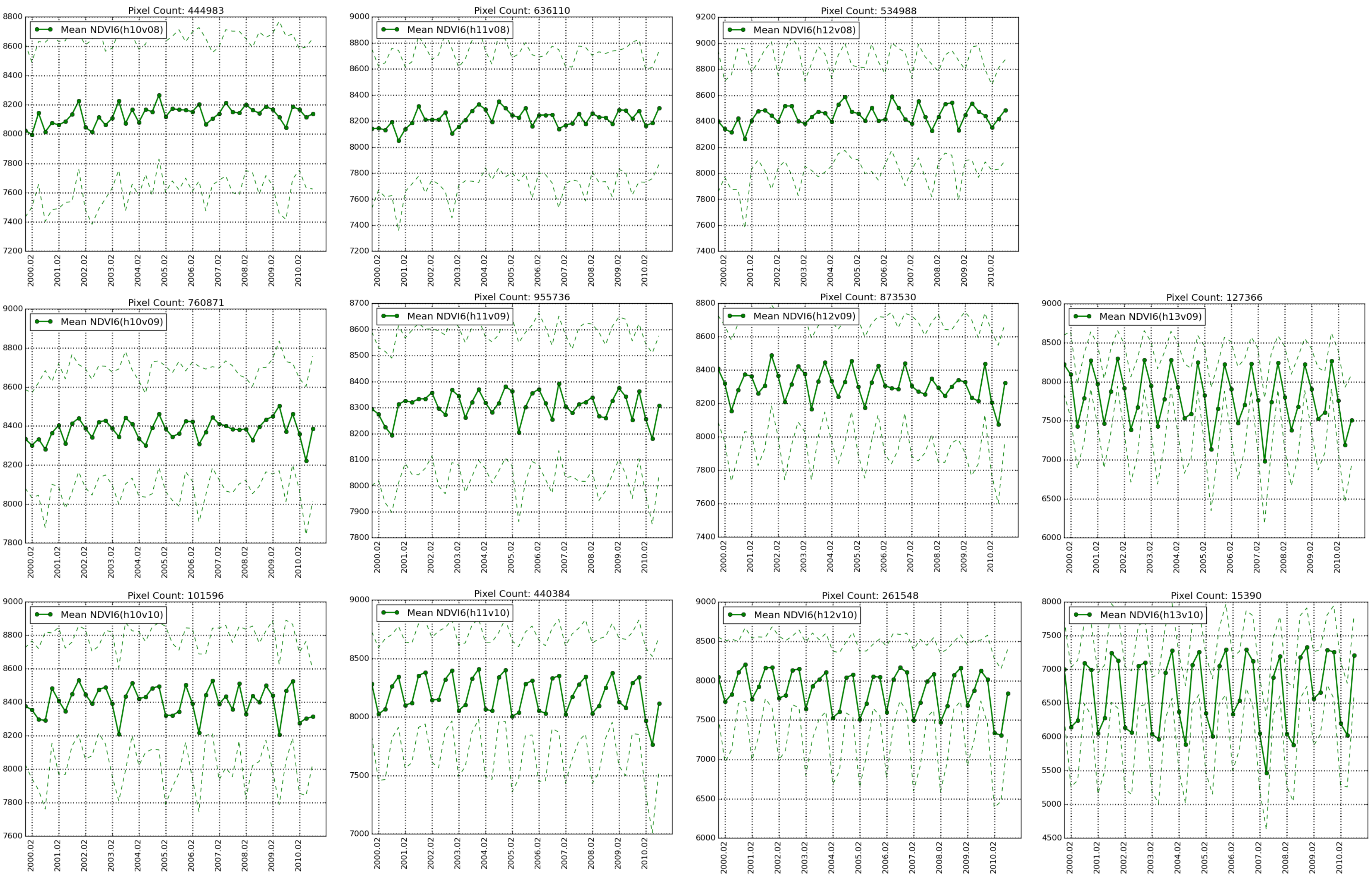
Data sets

- NDVI: Normalized Difference Vegetation Index (MOD13Q1: Terra MODIS, Collection 6)
  - NDVI = (NIR-RED)/(NIR+RED)
  - Spatial Resolution: 1km (originally 250m)
  - Temporal Resolution: seasonal (originally 16 day)
- Land Surface Temperature (day temperature) (MOD11A1: Terra MODIS, Collection 6)
  - Calculated using data from Thermal Infrared Region bands of MODIS sensor
  - Spatial Resolution: 1km (originally 1km)
  - Temporal Resolution: seasonal (originally daily)
- Precipitation (TRMM Product 3B43V7)
  - Total hydrometeor mass in the atmosphere column using microwave part of the spectrum
  - Spatial Resolution: 1km (originally 25km)
  - Temporal Resolution: seasonal (originally monthly)
- Elevation (GTOPO30)
  - Global digital elevation model
  - Spatial Resolution: 1 km
  - Derived feature: Slope (Horn's method)
- Landcover Mask (LCT from MOD12Q1.051, Collection 5)
  - Spatial Resolution: 1 km (originally 500m)

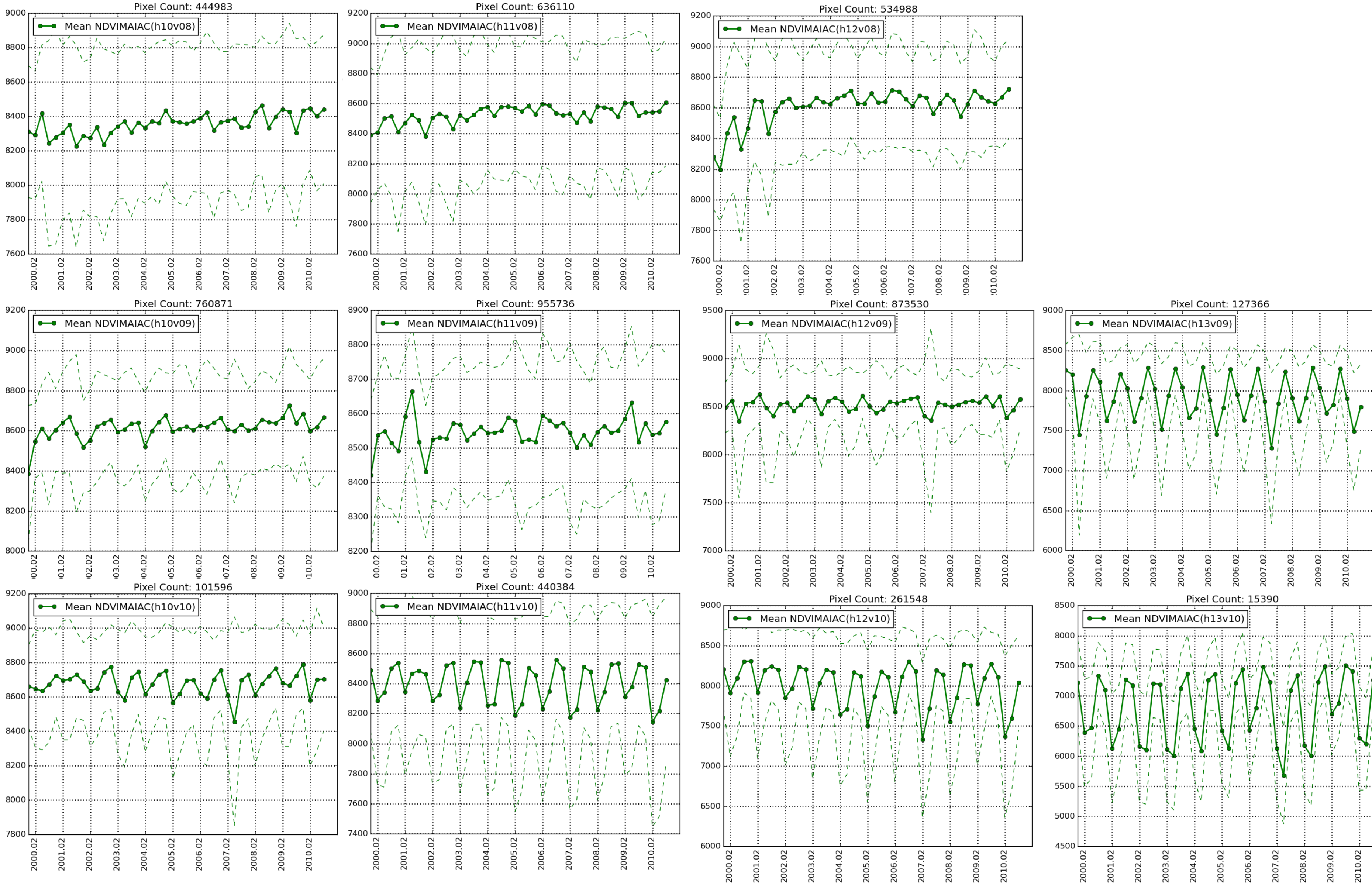
Amazon Region



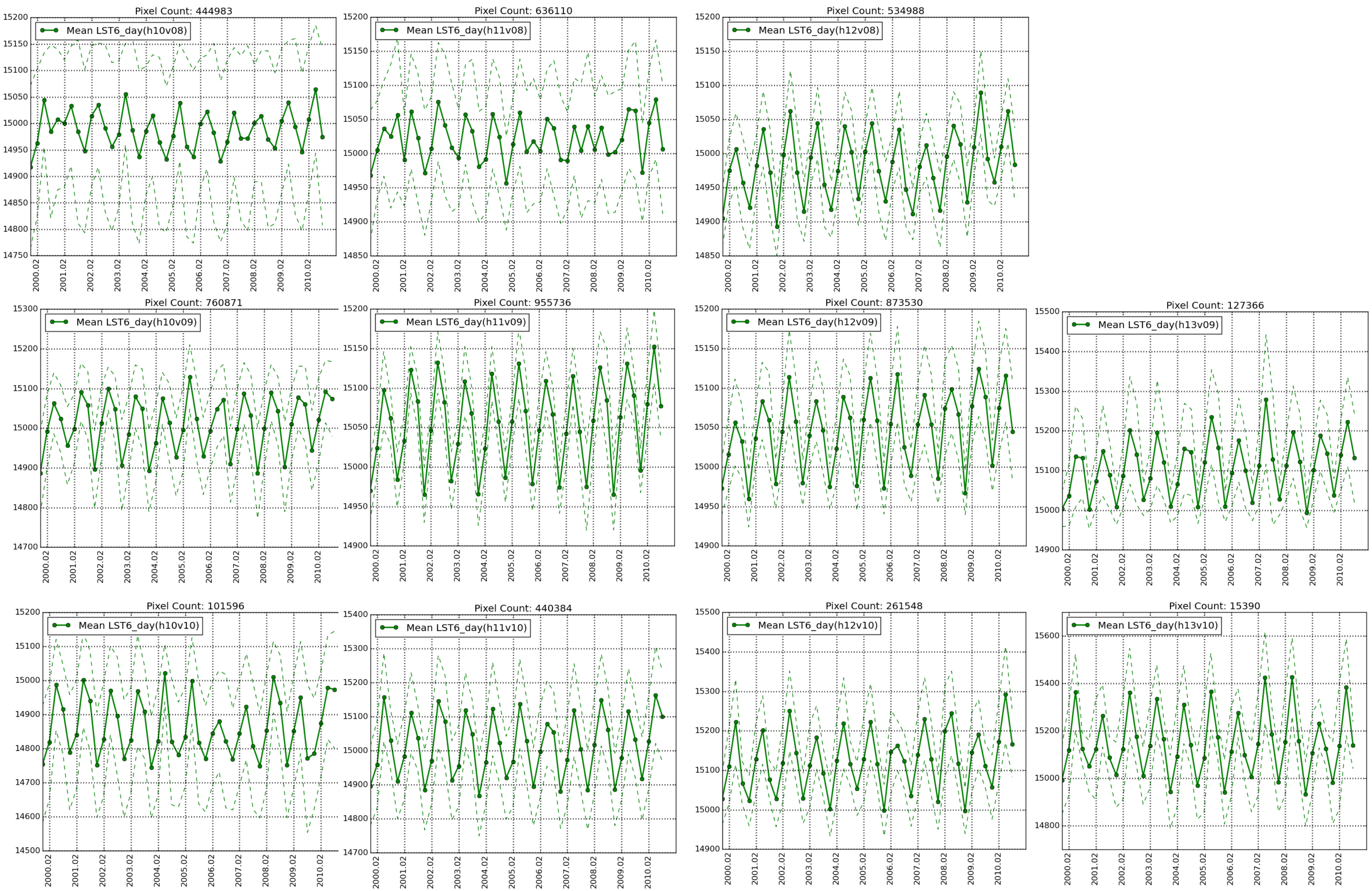
Trend analysis: MODIS NDVI



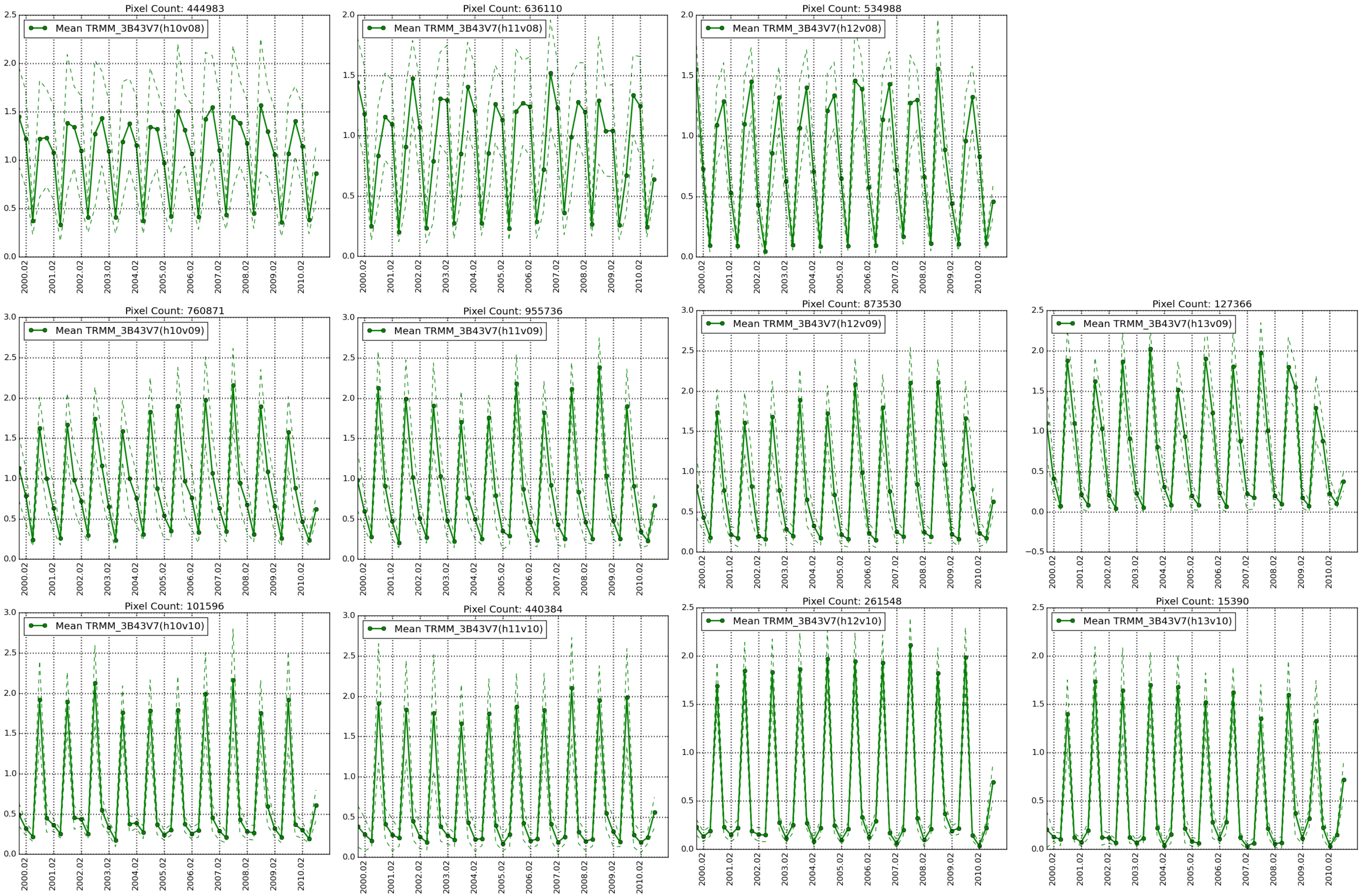
Trend analysis: MAIAC NDVI



Trend analysis: LST

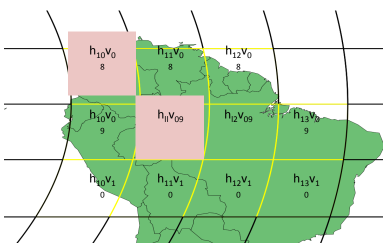


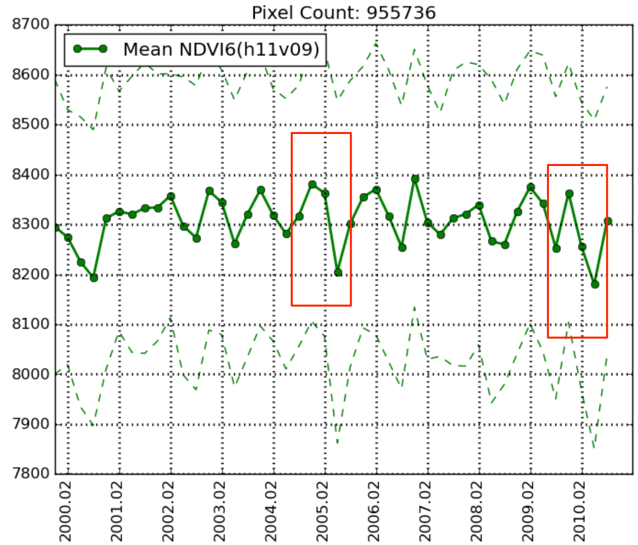
Trend analysis: TRMM

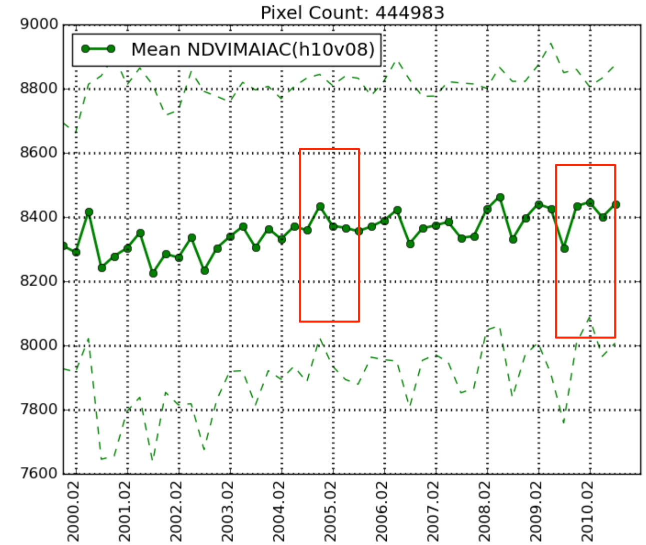


Drought signals

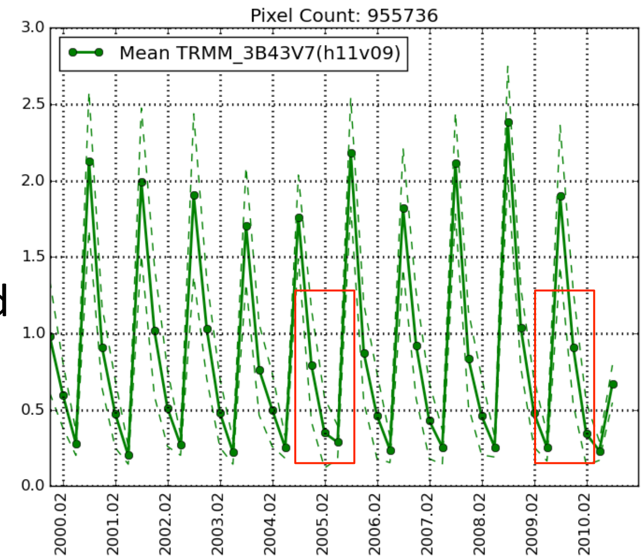
Vegetation trend

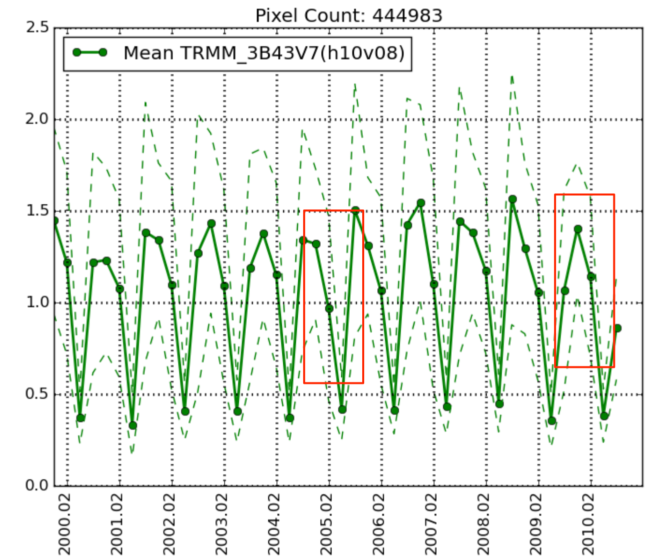






Precipitation trend





Regression setup

- Dependent variable (target): vegetation
- Independent variables (covariates/regressors)
  - Precipitation
  - Land surface temperature
  - Slope
  - Historical vegetation
- Modeling scenario: vegetation during the dry season as a function of historical (up to 1 year) and current covariates
- Learning model on data from 2003-2005
- Season definitions:
  - Dry season: July-September
  - Wet season: November – February
  - Dry to wet transition: October
  - Wet to dry transition: March – June
- Optimization metric: Normalized mean square error

Symbolic regression

- GP Settings
  - Age Fitness Pareto Algorithm
    - Objective: NMSE on training data, age of individual, size
    - 1 random individual added each generation
    - Pareto tournament selection of size p=2
  - 500 individuals evolved for 1000 generations
  - Function set: {+*-/log exp}
  - Terminals: independent variables and random constants drawn from a Gaussian distribution
- Final Solution
  - Choose point from Pareto front that is closest to ideal solution

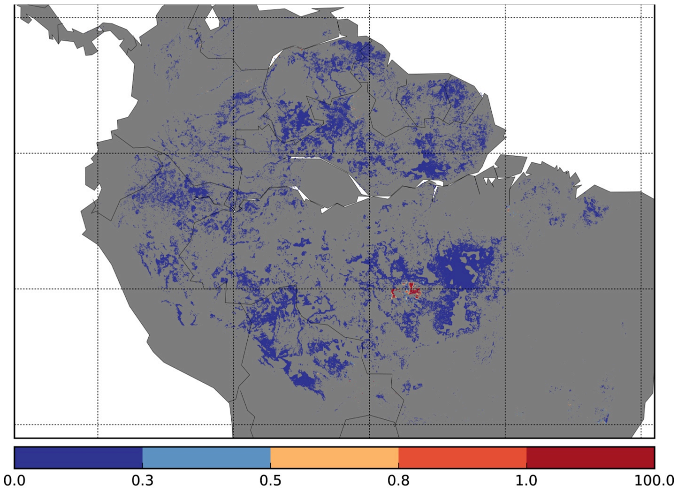
Results

- Training model on data from years 2003-2005
  - Symbolic regression based genetic programing
- Validation error on best model: 0.034
- Task using model:
  - Predicting vegetation for the dry season for entire Amazon for years 2006-2010

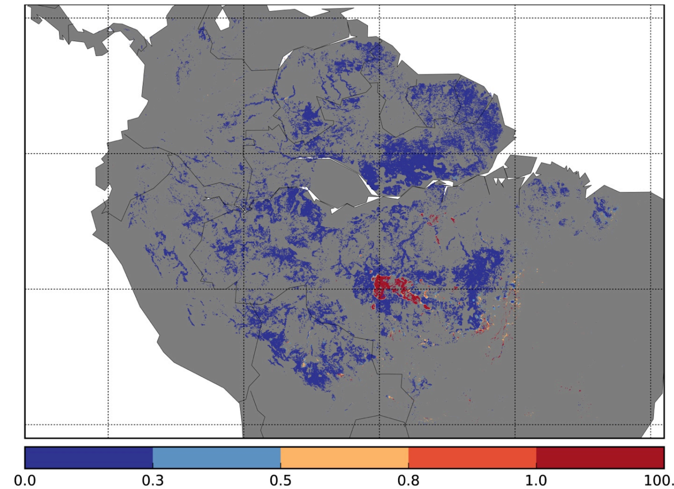
Prediction errors

Years	Tiles			
2006	0.034	0.022	0.011	0.025
	0.008	0.010	0.021	0.053
	0.017	0.016	0.080	0.108
2007	0.037	0.023	0.007	0.159
	0.016	0.013	0.334	0.142
	0.018	0.047	0.512	0.675
2008	0.028	0.019	0.008	0.053
	0.008	0.009	0.087	0.028
	0.024	0.016	0.153	0.068
2009	0.150	0.088	0.023	0.072
	0.023	0.023	0.145	0.153
	0.016	0.024	0.152	0.105
2010	0.028	0.018	0.007	0.079
	0.022	0.018	0.181	0.036
	0.046	0.051	0.092	0.373

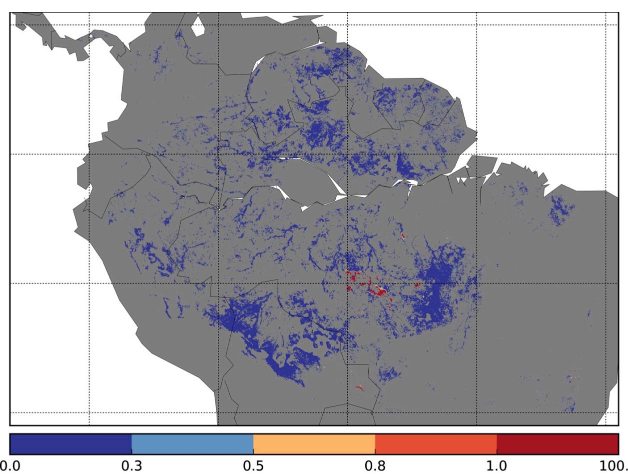
Spatial distribution of errors



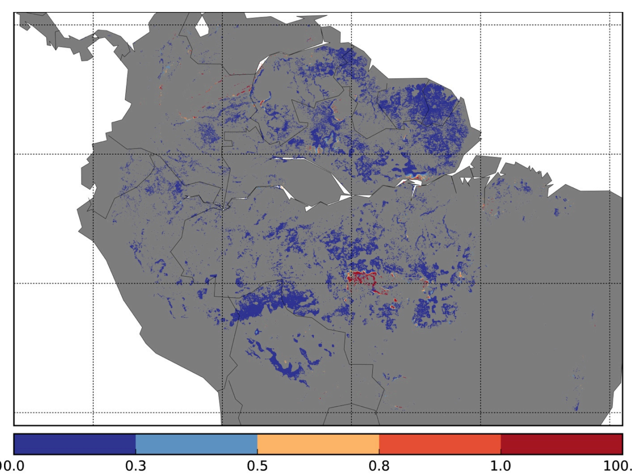
2006



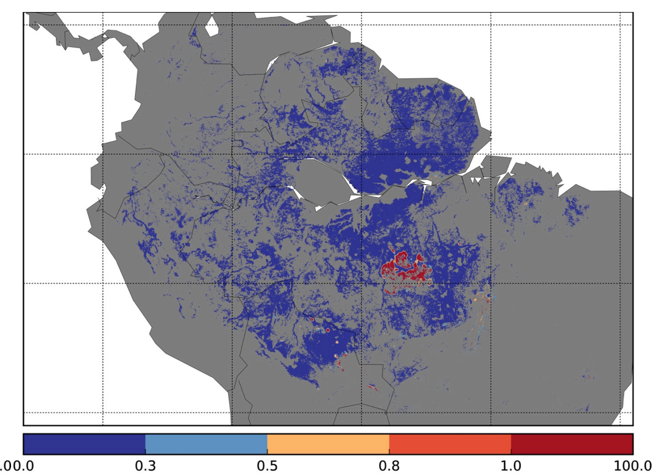
2007



2008

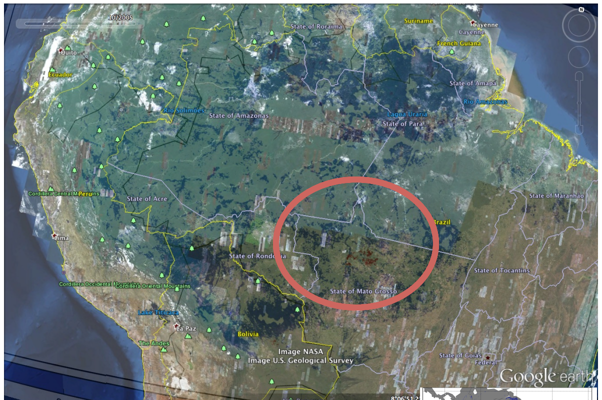
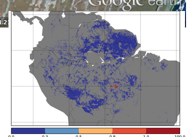


2009

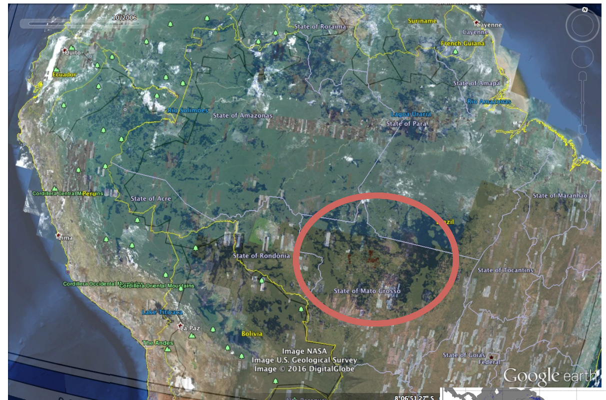
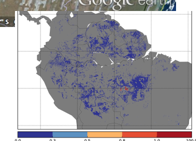


2010

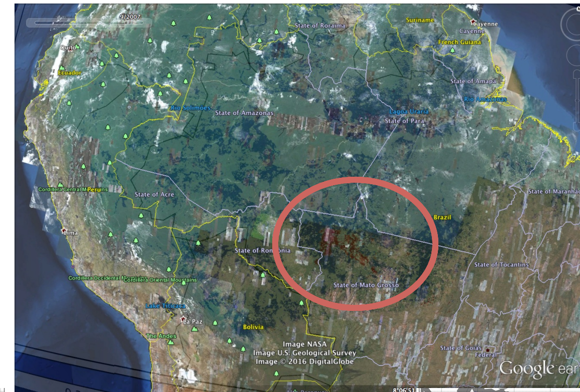
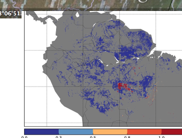
Explaining the high error regions

2005

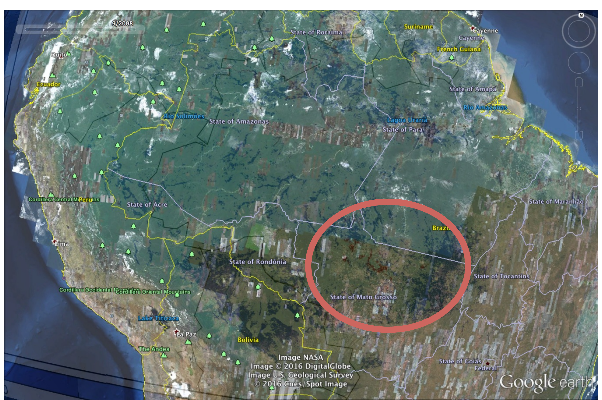
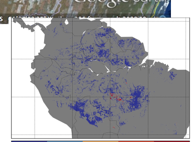



2006

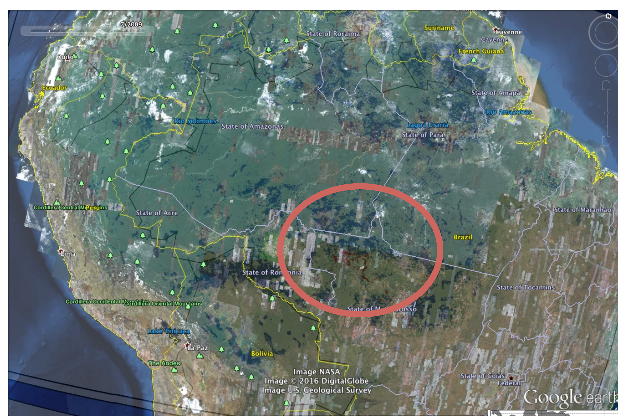
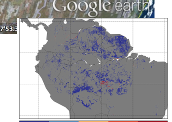



2007

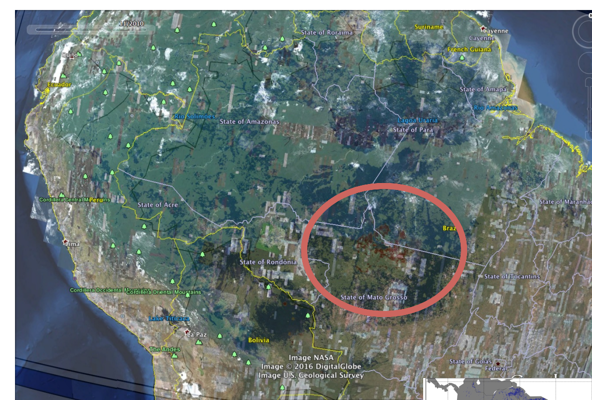
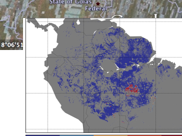
Urbanization increased in South-East region of Amazon from 2005 -> 2010

2008

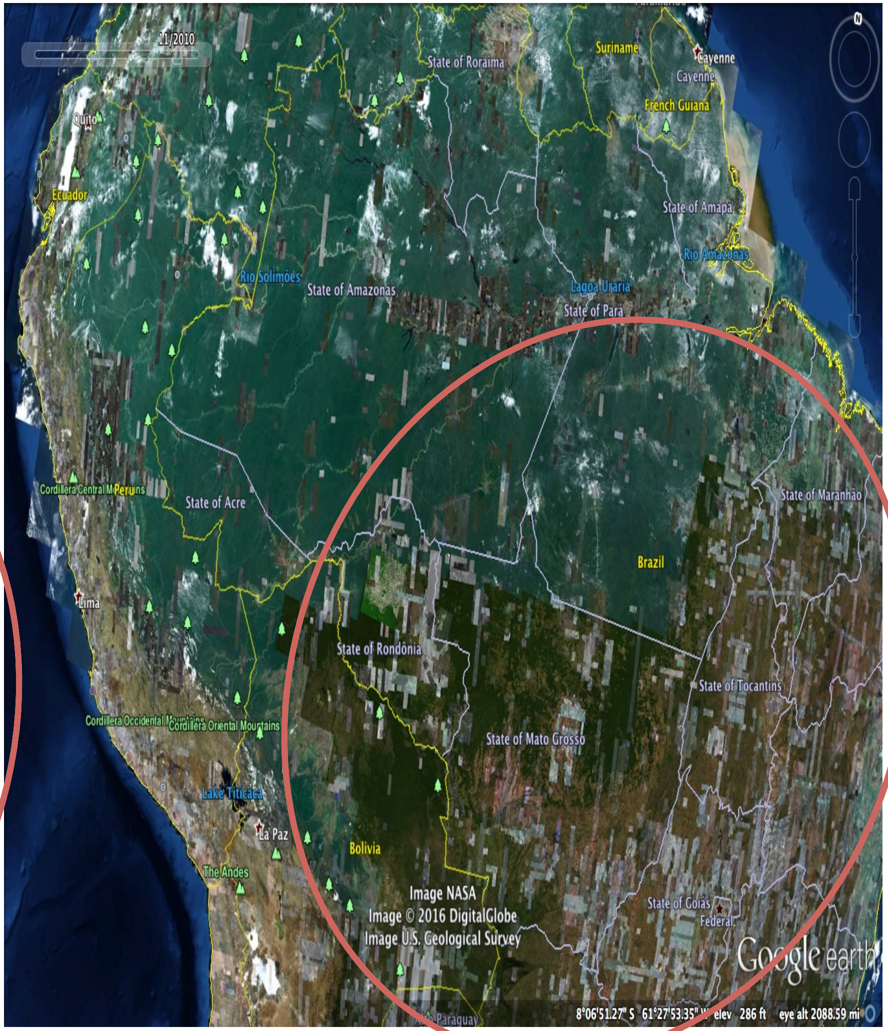
2009

2010

High error regions: logging

2003

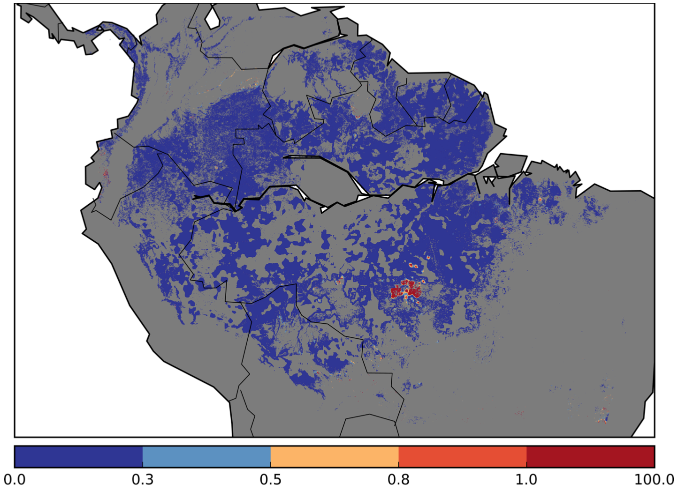


2010

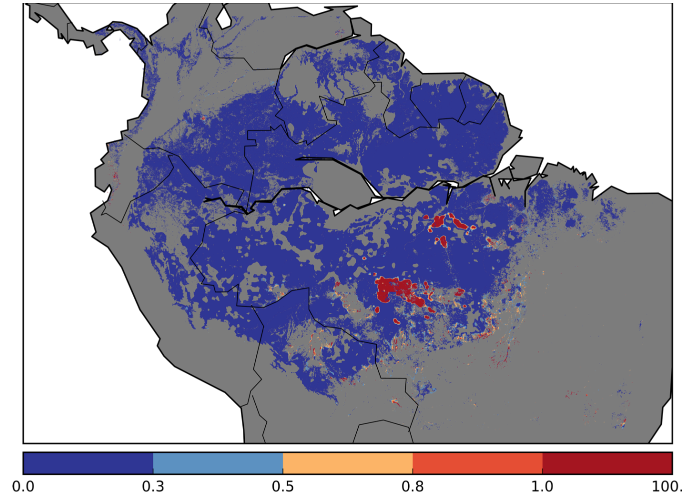
Model analysis

- Does this mean we have a model defining the exact relationship between precipitation, land surface temperature, and vegetation?
  - Can any data source for temperature and precipitation be used as regressors?
    - Use CRU data for verification:
      - Transfer learning for extending model to CRU data
      - New model

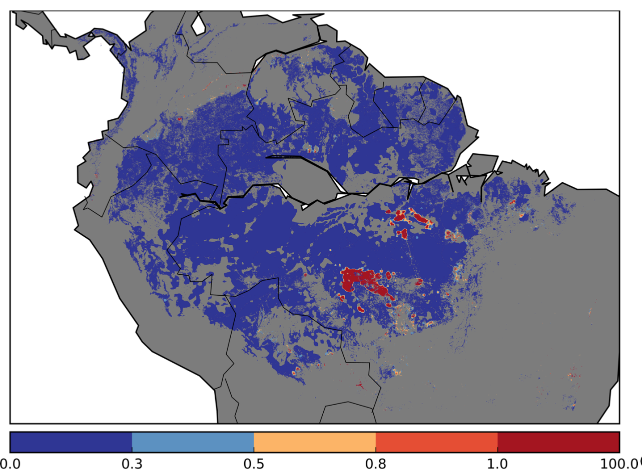
Spatial distribution of errors on CRU



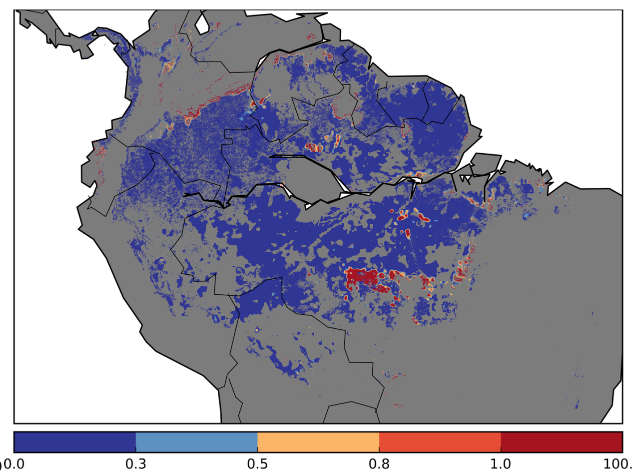
2006



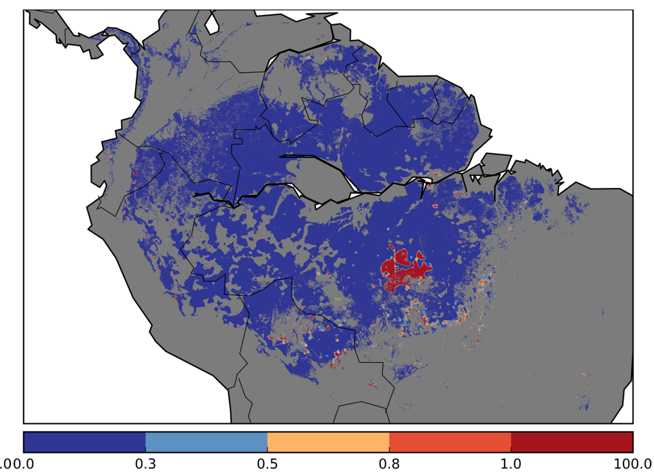
2007



2008



2009

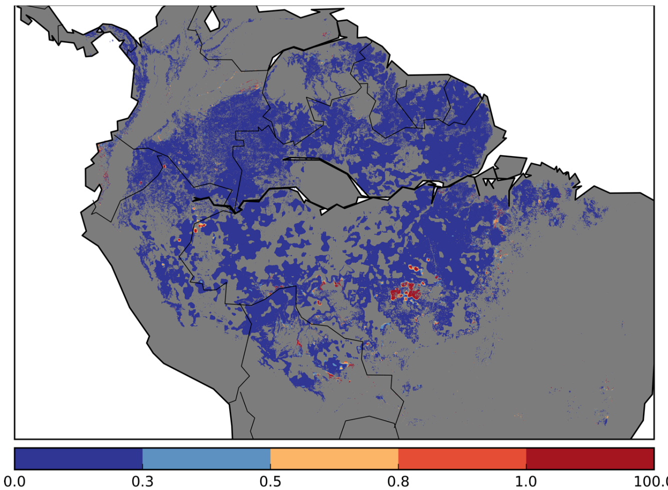


2010

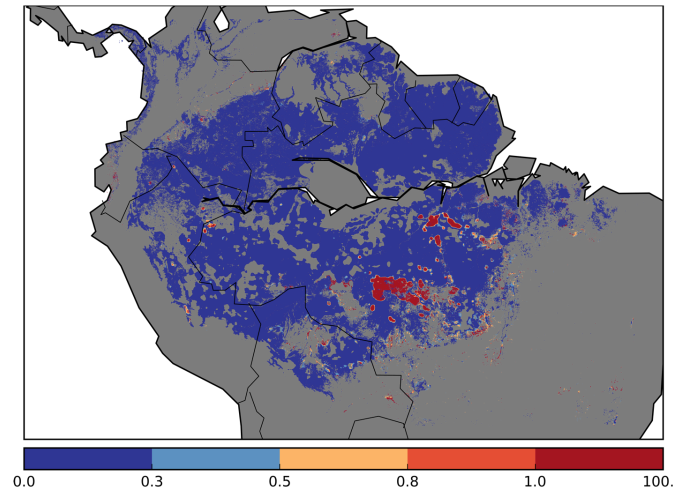
Predicting the future of Amazons

- Model built on observation data: not available for future prediction
  - Use global climate models' output as surrogate
  - NASA Earth Exchange Downscaled Climate Projections (NEX-DCP30)
    - Ensemble statistics of 33 models and 4 greenhouse emission scenarios from CMIP5
    - Temporal resolution: monthly
    - Spatial resolution: 25km
  - GP model transfer-learned for GCM data

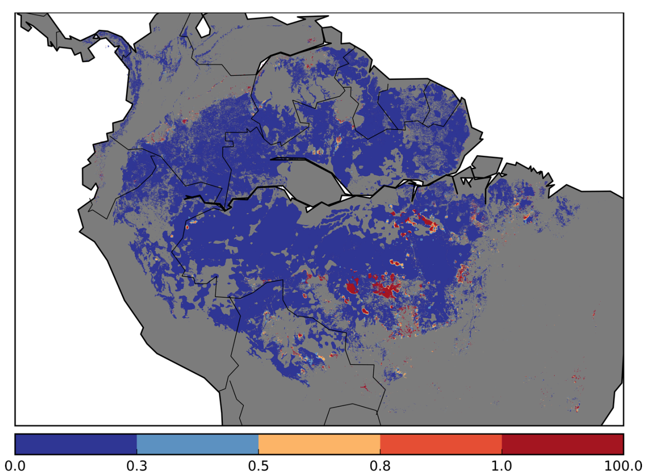
Predicting the future of Amazons



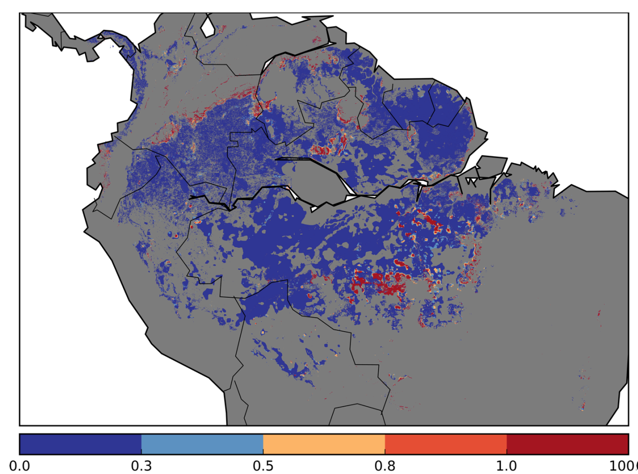
2006



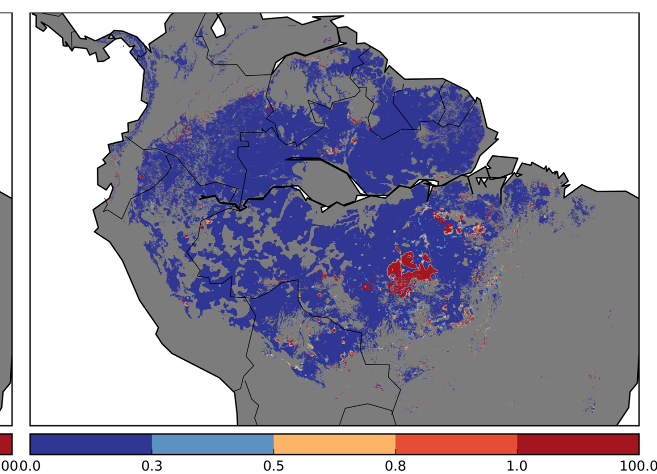
2007



2008



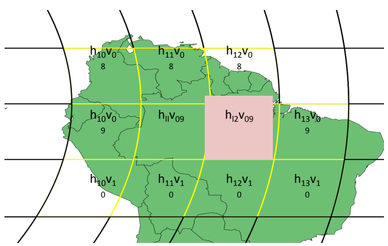
2009



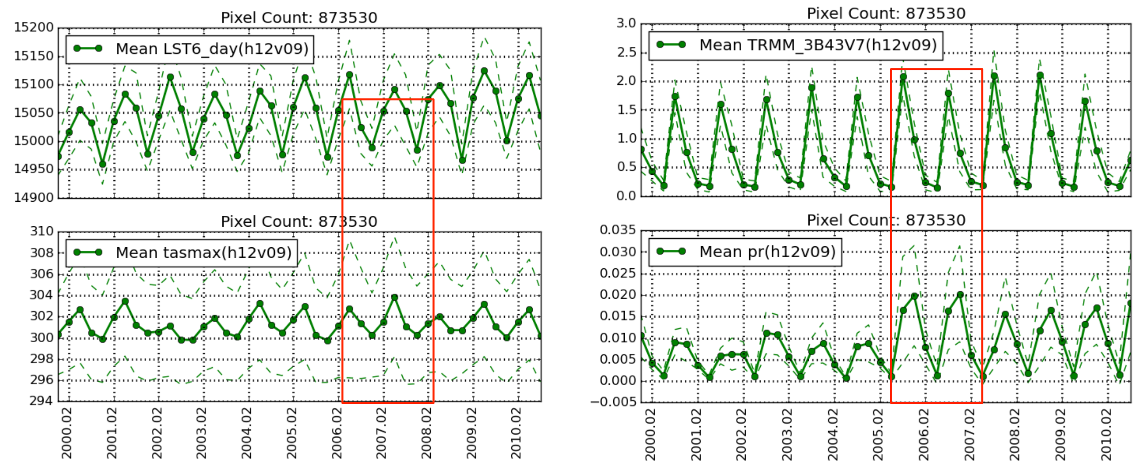
2010

Predicting the future of Amazons

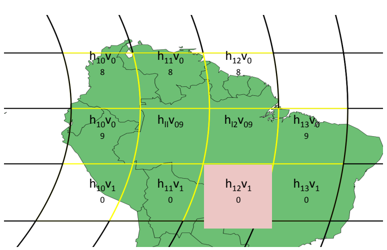
- Explaining high error regions



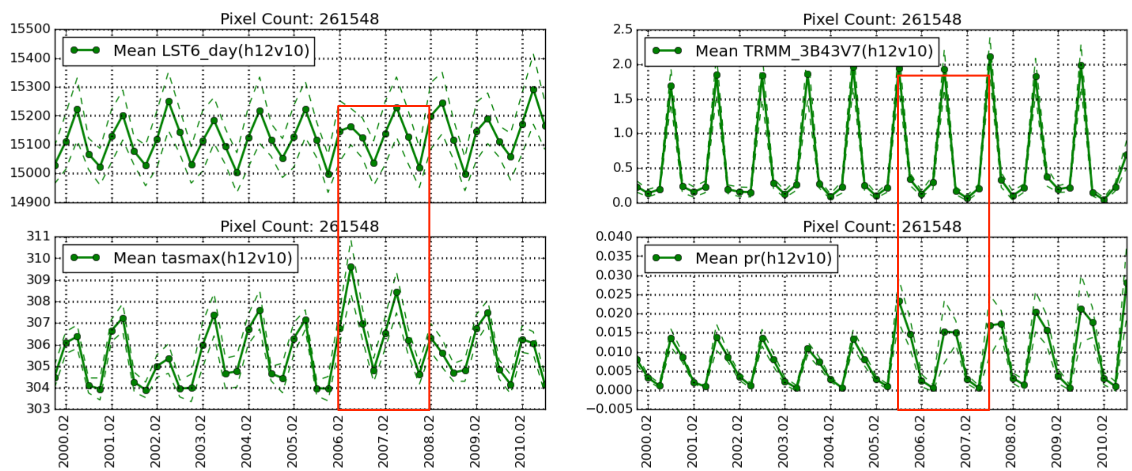
2007



Predicted (GCM) & observed (TRMM/LST) temperature (left) and precipitation (right)



2009



Conclusions and future directions

- Linear modeling does not capture the real dynamics between different climate variables and vegetation in the Amazon
- Symbolic regression uniquely identifies the feature space on which linear models can be learnt for different data sources
- Error distribution on Amazon NDVI indicates that model can be used to predict the future of Amazon rainforests, even under multiple severe droughts.

- Work in progress
  - Scaling model to global scale
  - Identifying latent factors in areas where model does not generalize
  - Using hierarchical modeling for combining local and global models

Questions?