Big Data in Climate:
Opportunities and Challenges for Machine Learning and Data Mining

Vipin Kumar

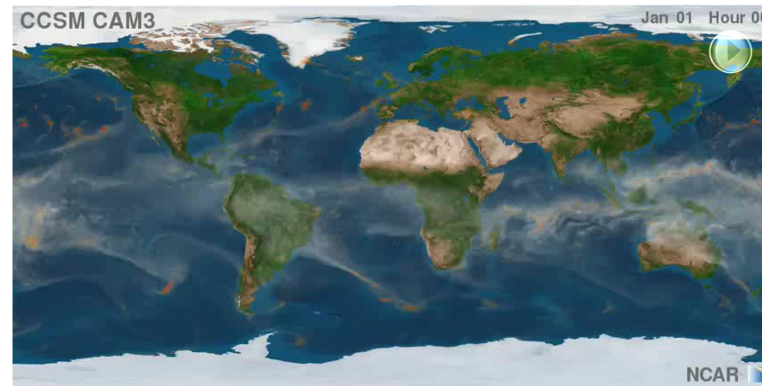
University of Minnesota

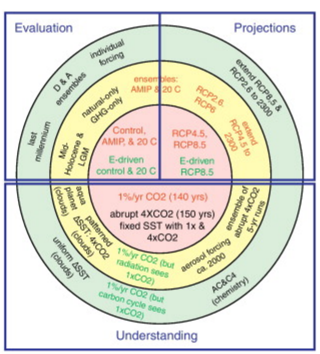
kumar@cs.umn.edu
www.cs.umn.edu/~kumar

Big Data in Climate

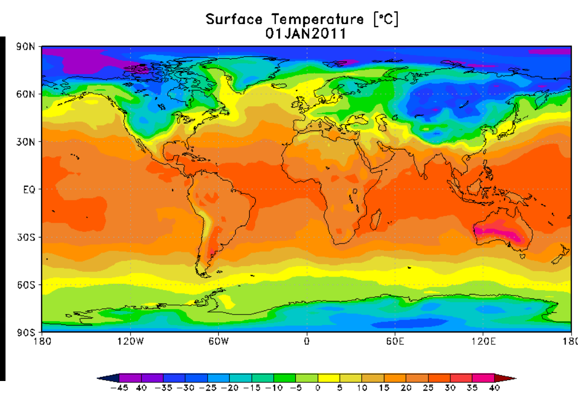
- Satellite Data
  - Spectral Reflectance
  - Elevation Models
  - Nighttime Lights
  - Aerosols
- Oceanographic Data
  - Temperature
  - Salinity
  - Circulation
- Climate Models
- Reanalysis Data
- River Discharge
- Agricultural Statistics
- Population Data
- Air Quality
- …

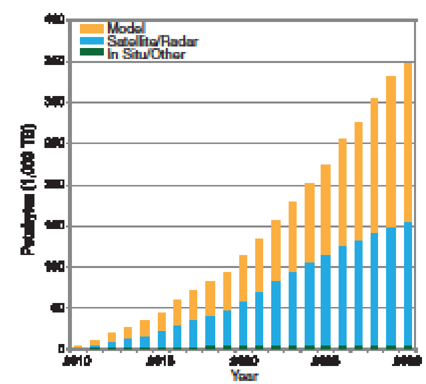
Source: NCAR





Source: NASA





Big Data in Climate

- Satellite Data
  - Spectral Reflectance
  - Elevation Models
  - Nighttime Lights
  - Aerosols
- Oceanographic Data
  - Temperature
  - Salinity
  - Circulation

- Climate Models
- Reanalysis Data

Source: NCAR

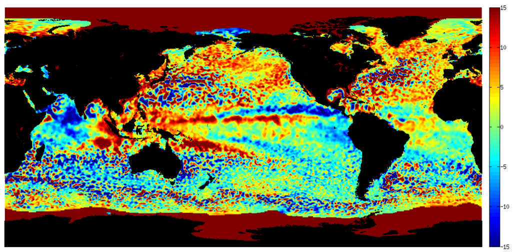
"Climate change research is now 'big science,' comparable in its magnitude, complexity, and societal importance to human genomics and bioinformatics."
(Nature Climate Change, Oct 2012)

Source: NASA

Five Year, $ 10m NSF Expeditions in Computing Project (1029711, PI: Vipin Kumar, U. Minnesota)

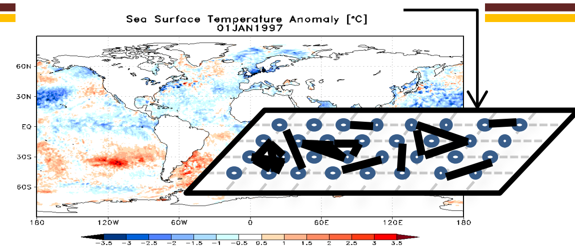
Understanding Climate Change: A Data-driven Approach

Research Highlights



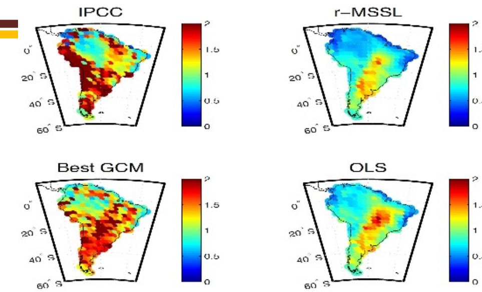
Pattern Mining: Monitoring Ocean Eddies

- Spatio-temporal pattern mining using novel multiple object tracking algorithms
- Created an open source data base of 20+ years of eddies and eddy tracks



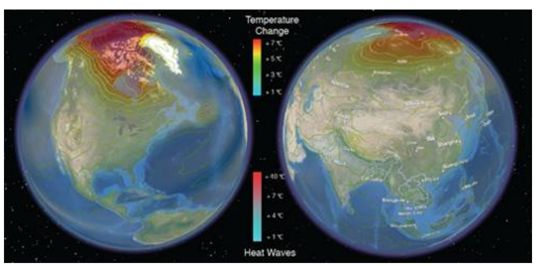
Network Analysis: Climate Teleconnections

- Scalable method for discovering related graph regions
- Discovery of novel climate teleconnections
- Also applicable in analyzing brain fMRI data



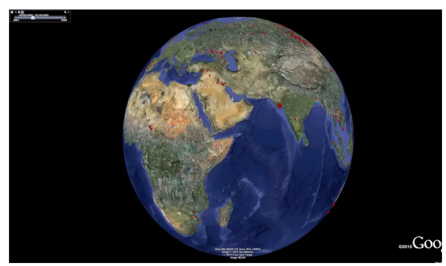
Sparse Predictive Modeling: Precipitation Downscaling

- Hierarchical sparse regression and multi-task learning with spatial smoothing
- Regional climate predictions from global observations



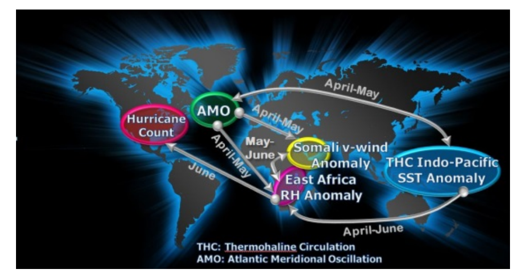
Extremes and Uncertainty: Heat waves, heavy rainfall

- Extreme value theory in space-time and dependence of extremes on covariates
- Spatiotemporal trends in extremes and physics-guided uncertainty quantification



Change Detection: Monitoring Ecosystem Distrubances

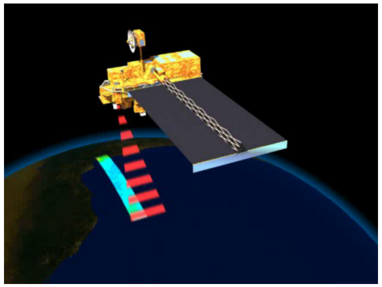
- Robust scoring techniques for identifying diverse changes in spatio-temporal data
- Created a comprehensive catalogue of global changes in surface water and vegetation, e.g. fires and deforestation.

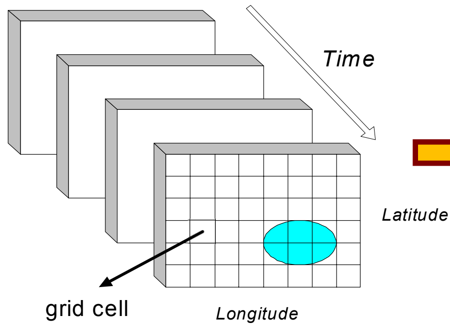


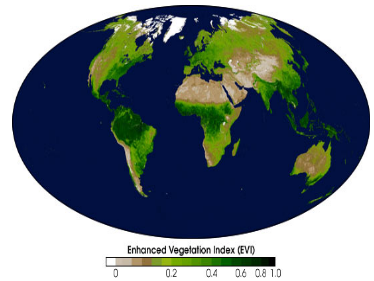
Relationship mining: Seasonal hurricane activity

- Statistical method for automatic inference of modulating networks
- Discovery of key factors and mechanisms modulating hurricane variability

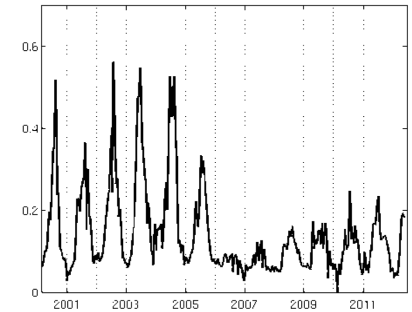
Big Data in Earth System Monitoring







A **vegetation index** measures the surface "greenness" – proxy for total biomass



This vegetation **time series** captures temporal dynamics around the site of the China National Convention Center

MODIS covers ~ 5 billion locations globally at 250m resolution daily since Feb 2000.

Data	Type	Coverage	Spatial Resolution	Temporal Resolution	Spectral Resolution	Duration	Availability
MODIS	Multispectral	Global	250 m	Daily	7	2000 - present	Public
LANDSAT	Multispectral	Global	30 m	16 days	7	1972 - present	Public
Hyperion	Hyperspectral	Regional	30 m	16 days	220	2001 - present	Private
Sentinal - 1	Radar	Global	5 m	12 days	-	2014 - present	Public
Quickbird	Multispectral	Global	2.16 m	2 to 12 days	4	2001 - 2014	Private
WorldView - 1	Panchromatic	Global	50 cm	6 days	1	2007 - present	Private

Monitoring Global Change: Case Studies

1. Global mapping of forest fires:
   - ❑ RAPT: Rare Class Prediction in Absence of Ground Truth
2. Global mapping of inland surface water dynamics
   - ❑ Heterogeneous Ensemble Learning and Physics-guided Labeling

Challenges

- Presence of noise, missing values, and poor-quality data
- Lack of representative ground truth
- High temporal variability
- Spatio-temporal auto-correlation
- Spatial and temporal heterogeneity
- Class imbalance (changes are rare events)
- Multi-resolution, multi-scale nature of data

Case Study 1:
Global Forest Fire Mapping

RAPT: Rare Class Prediction in Absence of True Labels

Global Forest Fires Mapping

Monitoring fires is important for climate change impact

A record number of more than 130 countries will sign the landmark agreement to tackle climate change at a ceremony at UN headquarters on 22 April, 2016.

“the best chance to save the one planet we have"


SEARCH

The New York Times

ENVIRONMENT

Delegates at Climate Talks Focus on Saving the World’s Forests

By JUSTIN GILLIS DEC. 10, 2015

The canopy of the forest in Puerto Viejo, Costa Rica, in October 2014. Climate change negotiations in Paris could lead to a sweeping effort to save the world's forests. Adriana Zehbrauskas for The New York Times

State-of-the-art: NASA MCD64A1

- Most extensively used global fire monitoring product
- Uses MODIS surface reflectance and Active Fire data in a predictive model
- Performance varies considerably across different geographical regions
- Known to have very low recall in tropical forests that play a critical role in regulating the Earth’s climate, maintaining biodiversity, and serving as carbon sinks

Predictive Modeling:

Traditional Paradigm

Given a feature vector $\boldsymbol{x} \in \mathbf{R}^d$
predict the class label $y \in \{0, 1\}$

Learn a classification function

$$f : \mathbf{R}^d \rightarrow \mathcal{Y}$$

which generalizes well on unseen data that comes from the same distribution as training data.

Explanatory Variable $\boldsymbol{x}_i \in \mathbf{R}^d$	**Target Label** $y_i \in \mathcal{Y} = \{0, 1\}$
$\boldsymbol{x}_1$	1
$\boldsymbol{x}_2$	0
$\boldsymbol{x}_3$	0
$\boldsymbol{x}_4$	1
.	.
$\boldsymbol{x}_N$	1

Predictive Modeling for Global Monitoring of Forest Fires

Challenges:

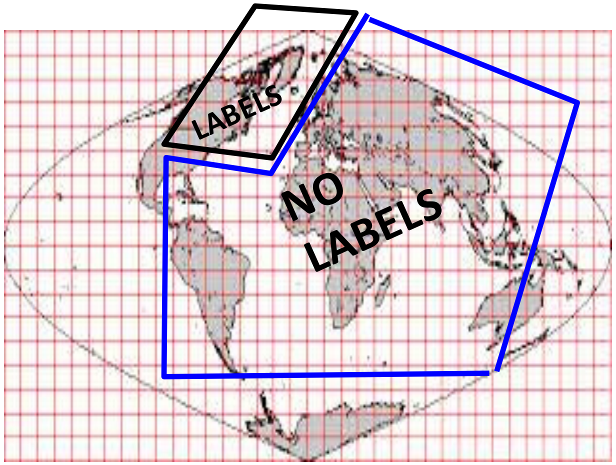
(1) Complete absence of target labels for supervision

(however, imperfect annotations of poor quality labels are available for every sample)

Variations in the relationship between the explanatory and target variable

- Geographical heterogeneity
- Seasonal heterogeneity
- Land class heterogeneity
- Temporal heterogeneity

$\boldsymbol{x}_i \in \mathbf{R}^d$	$y_i \in \mathcal{Y} = \{0, 1\}$
$\boldsymbol{x}_1$	?
$\boldsymbol{x}_2$	?
$\boldsymbol{x}_3$	?



Global availability of labeled samples for burned area classification

Predictive Modeling for Global Monitoring of Forest Fires

Challenges:

(1) Complete absence of target labels for supervision

(however, imperfect annotations of poor quality labels are available for every sample)

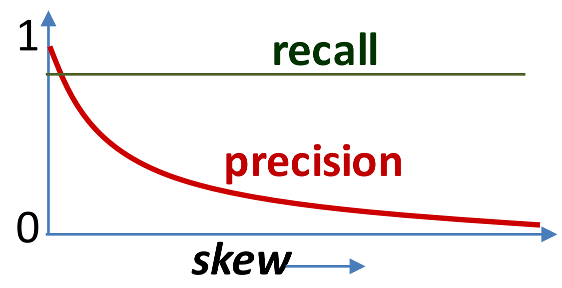
(2) Highly imbalanced classes

For eg. **California State**

Year 2008 (experienced maximum fire activity in last decade)

1,000 sq. km. of forests burned
out of a total
1,000,000 sq. km. forested area

True Positive Rate = 0.9
False Positive Rate = 0.01



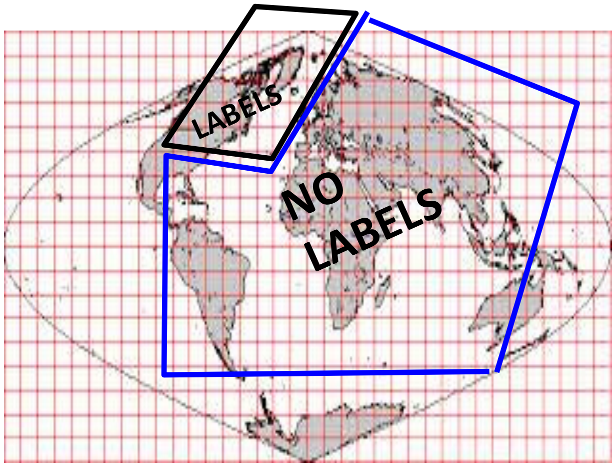
Predictive Modeling for Global Monitoring of Forest Fires

Challenges:

(1) Complete absence of target labels for supervision
(however, imperfect annotations of poor quality labels are available for every sample)

(2) Highly imbalanced classes

(3) How to evaluate performance of a model using imperfect labels?



Global availability of labeled samples for burned area classification

Predictive Modeling for Fire Monitoring

Challenges:

(1) Complete absence of target labels for supervision
(however, imperfect annotations of poor quality labels are available for every sample)

(2) Highly imbalanced classes

(3) How to evaluate performance of a model using imperfect labels?

State-of-the-art: NASA MCD64A1

- Domain heuristics and hand-crafted rules to identify high quality training samples
- Well known to have poor performance in the tropical forests.

Our Approach: RAPT [1]

- Trains classifiers using imperfect labels
  - Under certain assumptions, performance is comparable to classifiers trained on expert-annotated samples.
- Combines information in classifier output and imperfect labels to jointly maximize precision and recall
- Automatically identifies regions of poor performance.

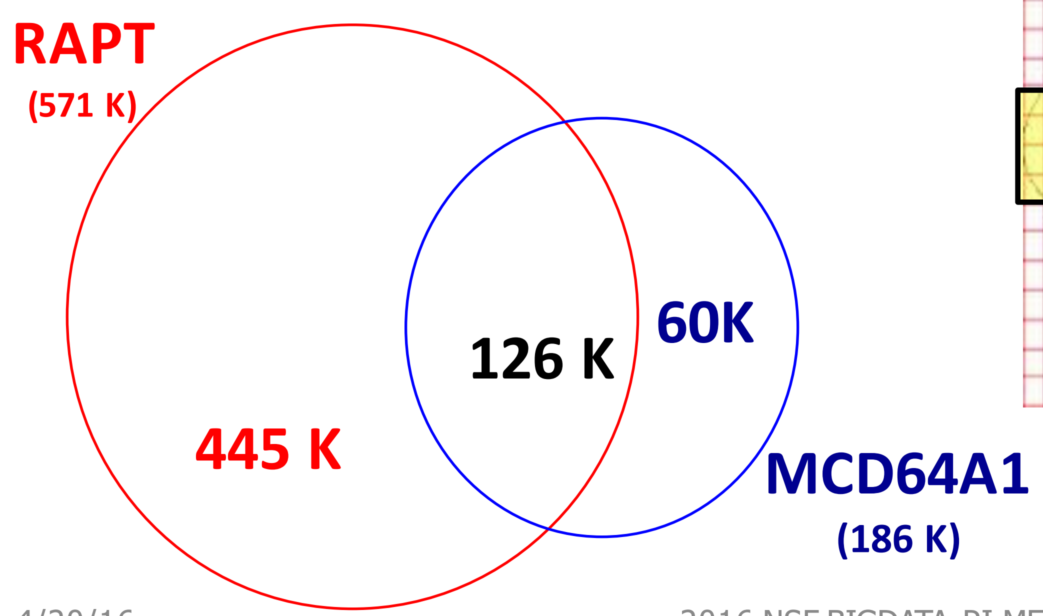
[1] Mithal (PhD Dissertation)

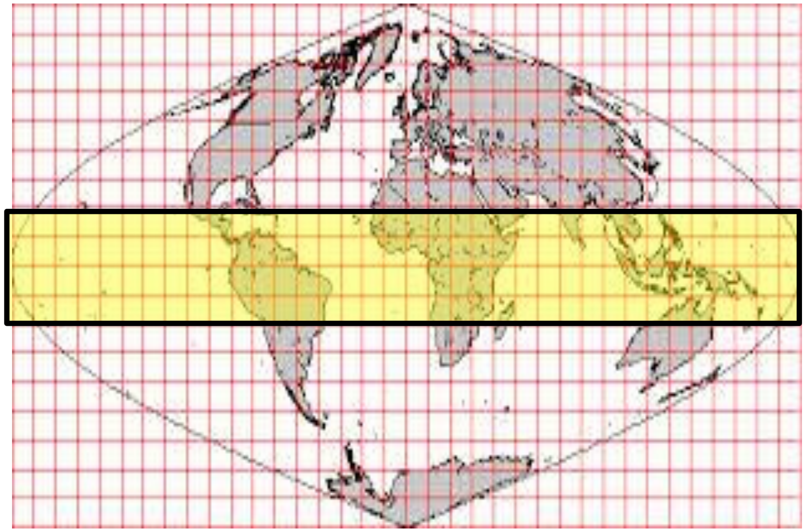
Global Monitoring of Fires in Tropical Forests

Fires in tropical forests during 2001-2014

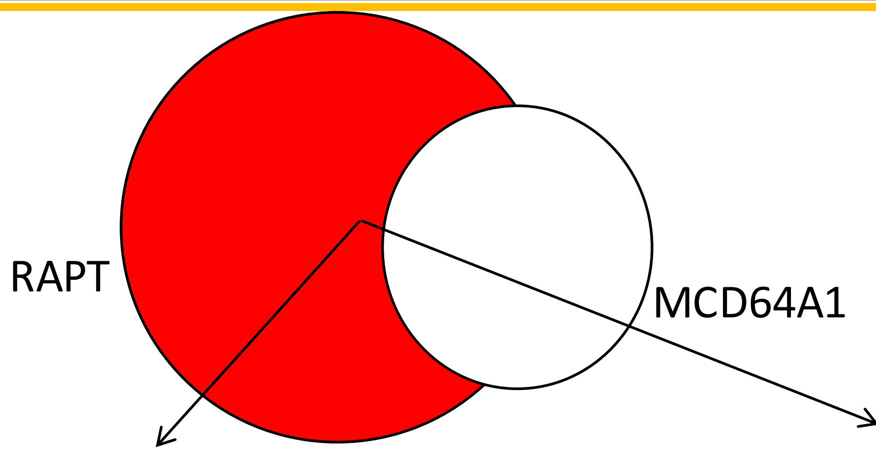
571 K sq. km. burned area found in tropical forests

- *more than three times the total area reported by state-of-art NASA product: MCD64A1*.





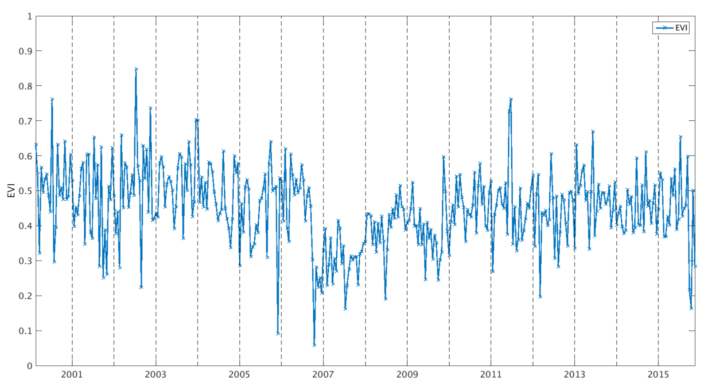
Validation

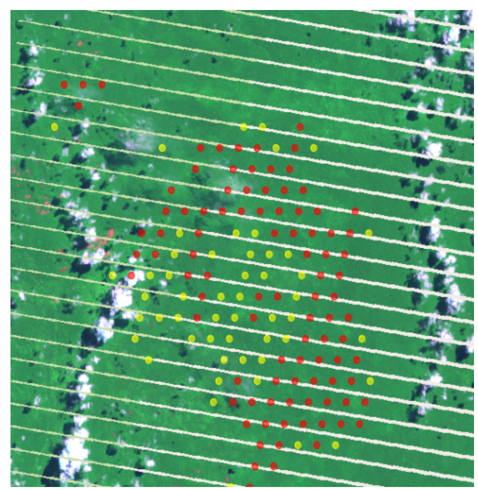


Multiple lines of evidence indicate that RAPT-only points are actual forest fires

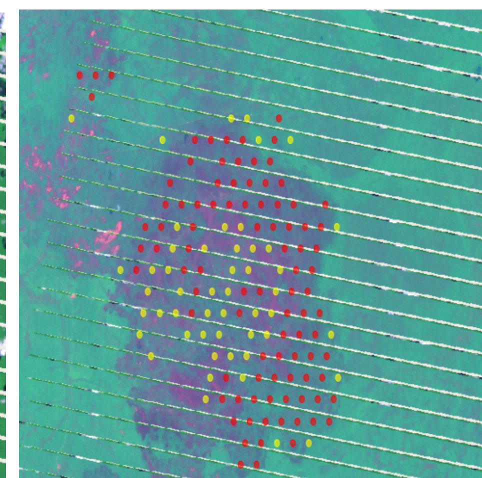
Burn scar in Landsat composite

Change in Vegetation series





Before Fire Event

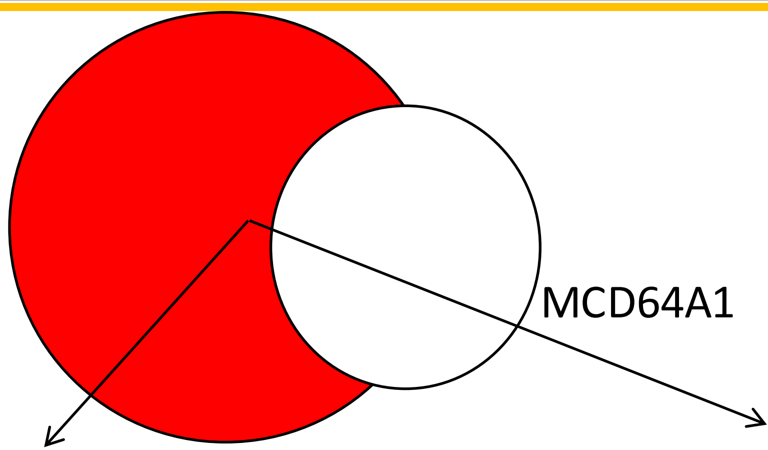


After Fire Event

Sudden drop followed by recovery is a key signature of forest fires

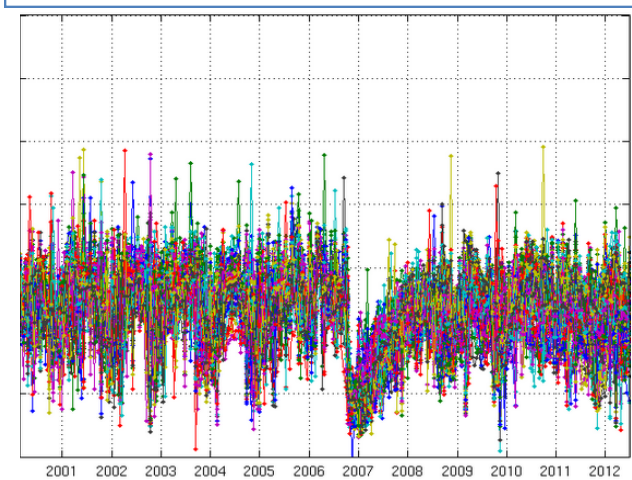
Landsat false-color composite shows the scar after the fire event identified by RAPT

Validation



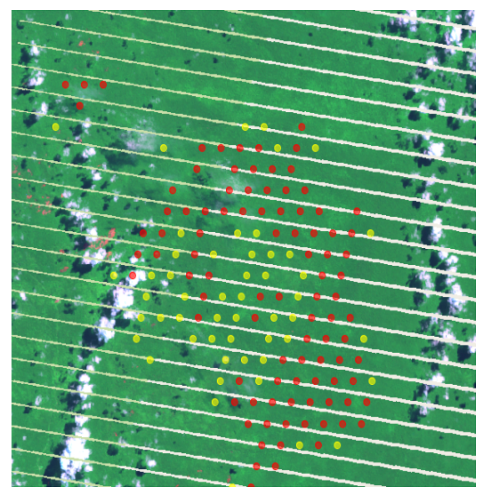
Multiple lines of evidence indicate that RAPT-only points are actual forest fires

Change in Vegetation series

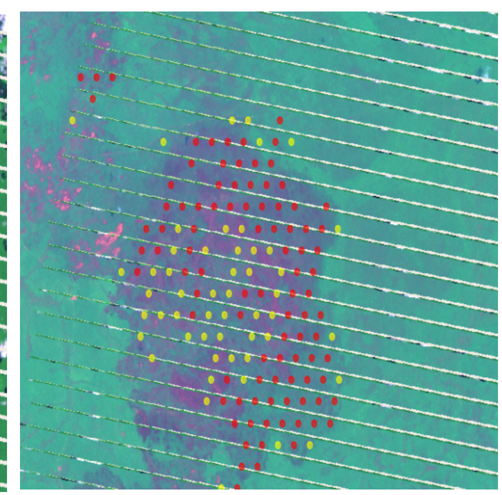


Synchronized drop followed by recovery is a key signature of forest fires

Burn scar in Landsat composite



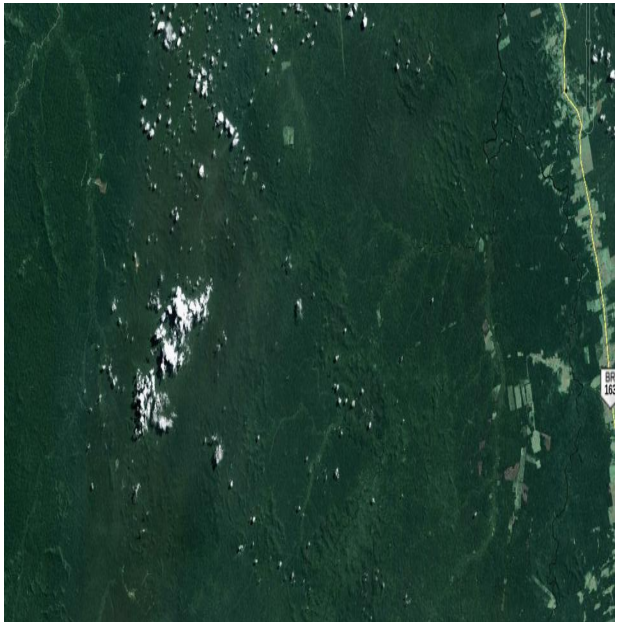
Before Fire Event



After Fire Event

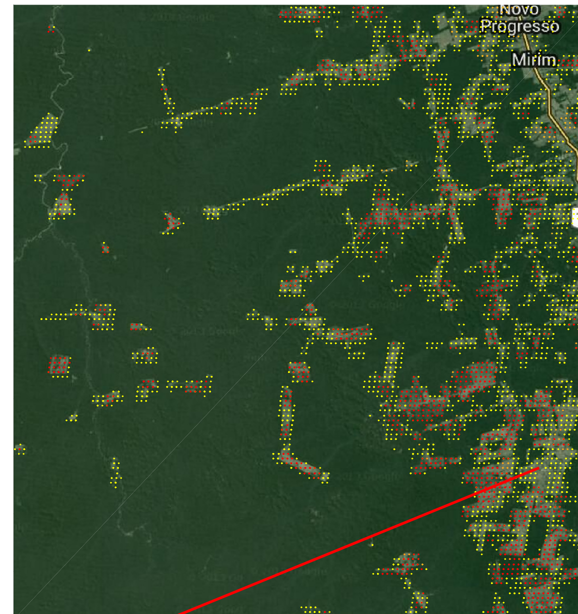
Landsat false-color composite shows the scar after the fire event identified by RAPT

Active Deforestation Fronts in Amazon



Google Earth Image: Year 2002

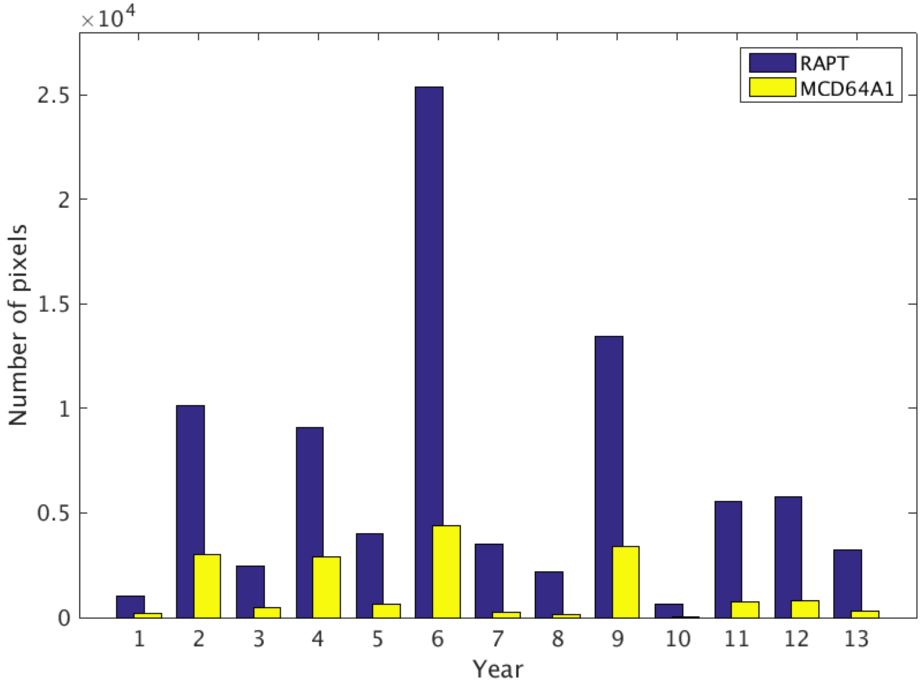
Google Earth Image: Year 2015



RAPT detection 2002-2014
(RAPT only, Common)

Burn Detection					B	B	B						
Land cover	F	F	F	F	F	F	F	N	N	N	N	N	N
Year	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014

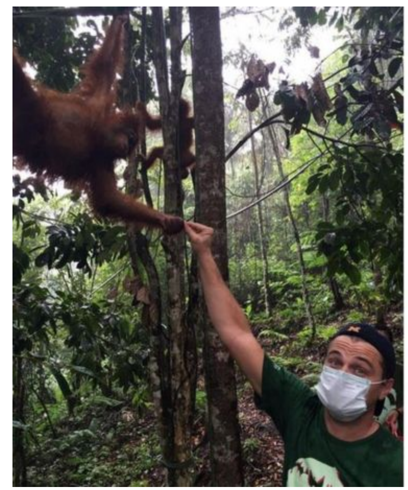
Palm Oil Plantations in Indonesia



Number of 500 m pixels in forests that were identified as burned and converted to plantations[1] in Indonesia from years 2001 to 2013.

[1]Plantation maps obtained from Global Forest Watch

Indonesia 'may blacklist Leonardo DiCaprio over palm oil comments'



"A world-class biodiversity hotspot... but palm oil expansion is destroying this unique place." – Leonardo DiCaprio

Case Study 2: Global Mapping of Surface Water Dynamics

Heterogeneous Ensemble Learning and Physics-guided Labeling

http://z.umn.edu/monitoringwater

Importance of Monitoring Global Surface Water Dynamics

Brazil's Severe Drought Dries Up Reservoirs
California is not alone: São Paulo is also facing severe water restrictions.

Oil-Rich Persian Gulf Looks to Renewables to Avert Water Crisis
BloombergBusiness January 19, 2016

Kariba Dam Water Levels 'Dire,' Zambian Energy Minister Says
January 8, 2016

nature International weekly journal of science

Published online 12 August 2009 | *Nature* **460**, 789 (2009) | doi:10.1038/460789a

News

Satellite data show Indian water stocks shrinking

Groundwater depletion raises spectre of shortages.

Effect Of Climate Change On Agriculture: Droughts, Heat Waves Cut Global Cereal Harvests By 10 Percent In 50 Years
TECH TIMES January 7

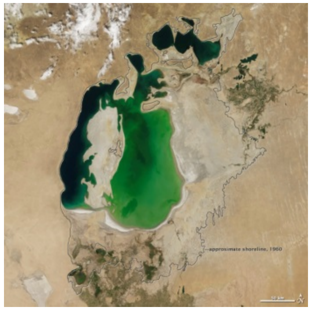
Smithsonian.com

The Colorado River Runs Dry
Dams, irrigation and now climate change have drastically reduced the once-mighty river. Is it a sign of things to come?


Cedo Caka Lake in Tibet, 1984


Cedo Caka Lake in Tibet, 2011

Melting of glacial lakes in Tibet


Aral Sea in 2000


Aral Sea in 2014

Shrinking of Aral Sea since 1960s

Importance of Monitoring Global Surface Water Dynamics

Brazil's Severe Drought Dries Up Reservoirs
California is not alone: São Paulo is also facing severe water restrictions.

Oil-Rich Persian Gulf Looks to Renewables to Avert W
Crisis BloombergBusiness J

Kariba Dam Water Levels

Effect Of Climate C
Droughts, Heat Wav
Harvests By 10 Per
TECH TIMES Jan

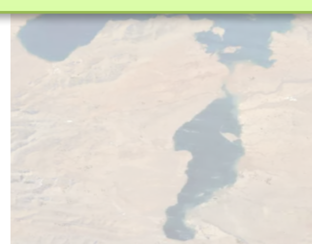
nature International weekly journal of science
Published online 12 August 2009 | *Nature* 460, 789 (2009) | doi:10.1038/460789a
News
w Indian water stocks
es spectre of shortages.

ian.com
ver Runs Dry
limate change have drastically reduced
a sign of things to come?

Opportunity in using Remote Sensing Data

- Multi-spectral data
  - MODIS (at 500m, from 2000)
  - Landsat (at 30m, from 1970s)
- Can be used to classify every location at a given time as water or land (binary classes)
- Ground truth on specific dates available from various sources: SRTM, GLWD

Cedo Caka Lake in Tibet, 1984



Cedo Caka Lake in Tibet, 2011

Melting of glacial lakes in Tibet

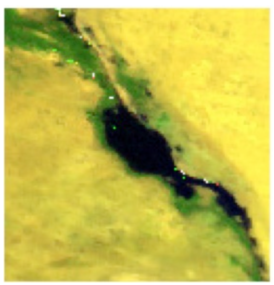
Aral Sea in 2000

Aral Sea in 2014

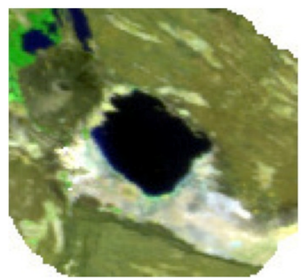
Shrinking of Aral Sea since 1960s

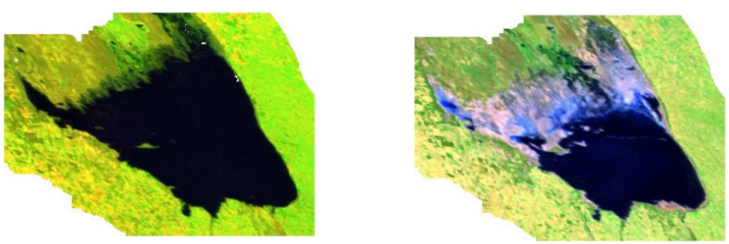
Challenges for Traditional Big Data Methods in Monitoring Water

- **Challenge 1: Heterogeneity in space and time**
  - Water and land bodies look different in different regions of the world
  - Same water body can look different at different time-instances

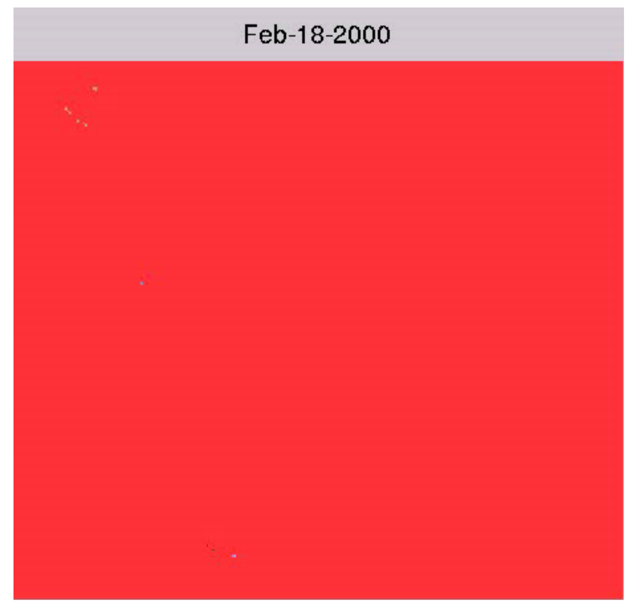

Great Bitter Lake, Egypt


Lake Tana, Ethiopia


Lake Abbe, Africa


Mar Chiquita Lake, Argentina in 2000 (left) and 2012 (right)

- **Challenge 2: Data Quality**
  - Noise: clouds, shadows, atmospheric disturbances
  - Missing data



Poyang Lake, China
(Pink color shows missing data)

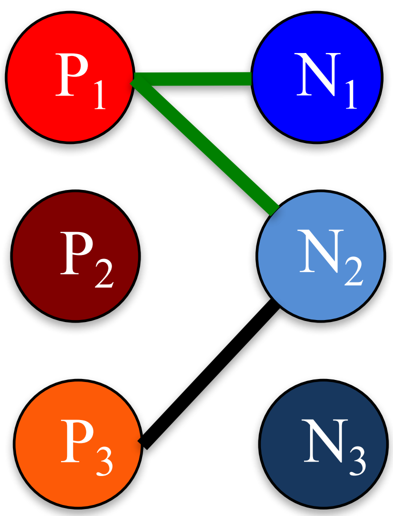
Method Innovations for Monitoring Water

- **Ensemble Learning Methods for Handling Heterogeneity in Data** [1,2]

Learn an ensemble of classifiers to distinguish b/w different pairs of positive and negative modes

Positive Modes (Water)

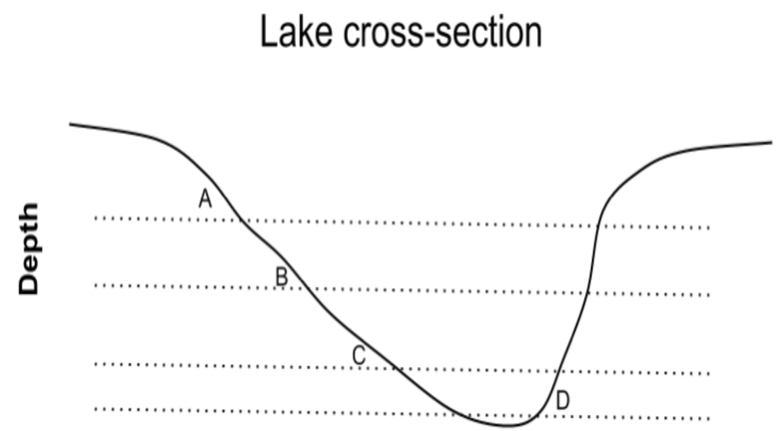
Negative Modes (Land)



[1] Karpatne et al. SDM 2015
[2] Karpatne et al. ICDM 2015

- **Using Physics Guided Labeling to Handle Poor Data Quality** [3,4]

Use elevation information to constrain physically-consistent labels



Elevation A > B > C > D

[3] Khandelwal et al. ICDM 2015
[4] Mithal (PhD Dissertation)

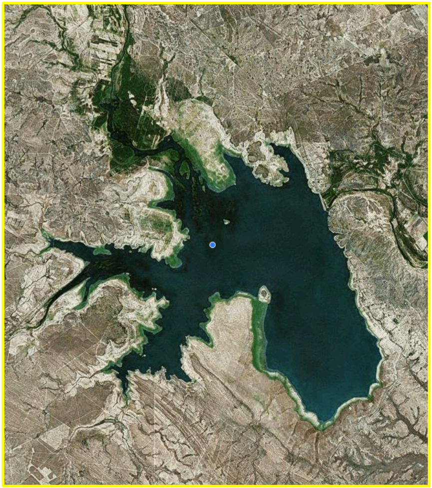
A Global Water Monitoring System
http://z.umn.edu/monitoringwater

- Summary of Capabilities:
  - Maps the dynamics of all major water bodies (surface area > 2.5 km^2) in the last 15 years across the world
  - Finds changes in river morphology (river meandering, delta erosion)
  - Detects the construction of new dams and reservoirs
  - Demonstrates strong relationships b/w surface water and ground water detected by GRACE

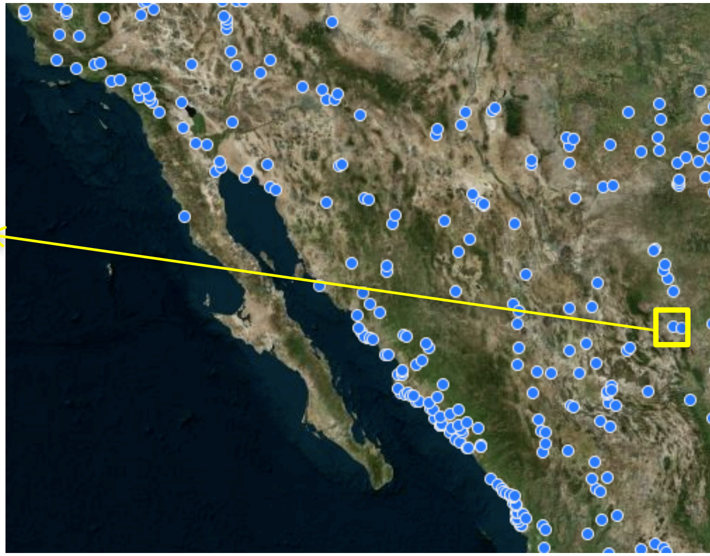
Global Maps of Water Bodies

Every blue dot is a water body, present in the last 15 years, with size greater than 2.5 km^2

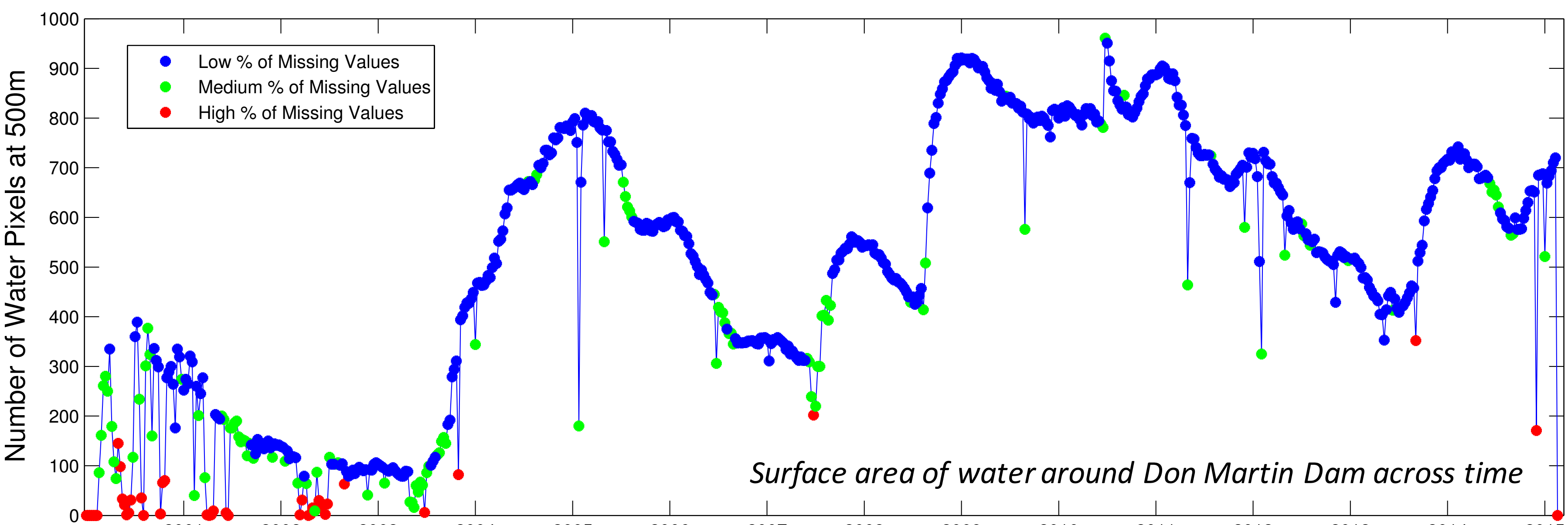
Showing Surface Water Dynamics



Don Martin Dam, Mexico



Annual Landsat Time-lapse of this region
(Courtesy: Google Earth Engine)

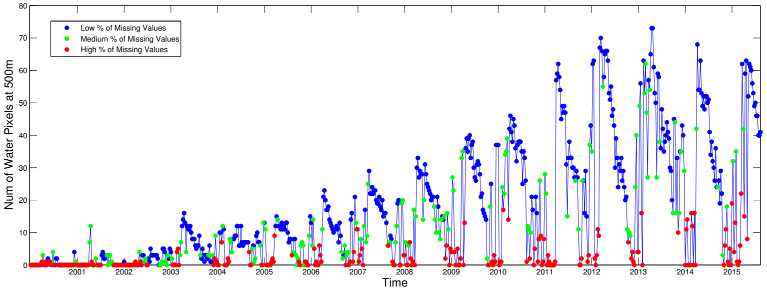


Surface area of water around Don Martin Dam across time

Regions of Change in South America

Red Dots *(Water Gain):*

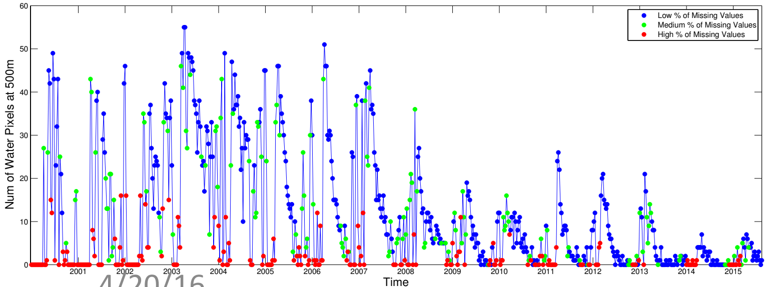
Region of size > 2.5 km^2 that have changed from land to water in the last 15 years



Example time series of a *Water Gain* region

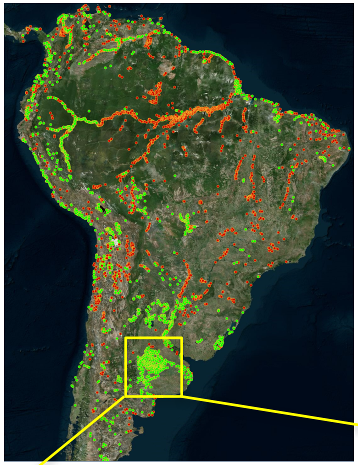
Green Dots (Water Loss):

Region of size > 2.5 km^2 that have changed from water to land in the last 15 years

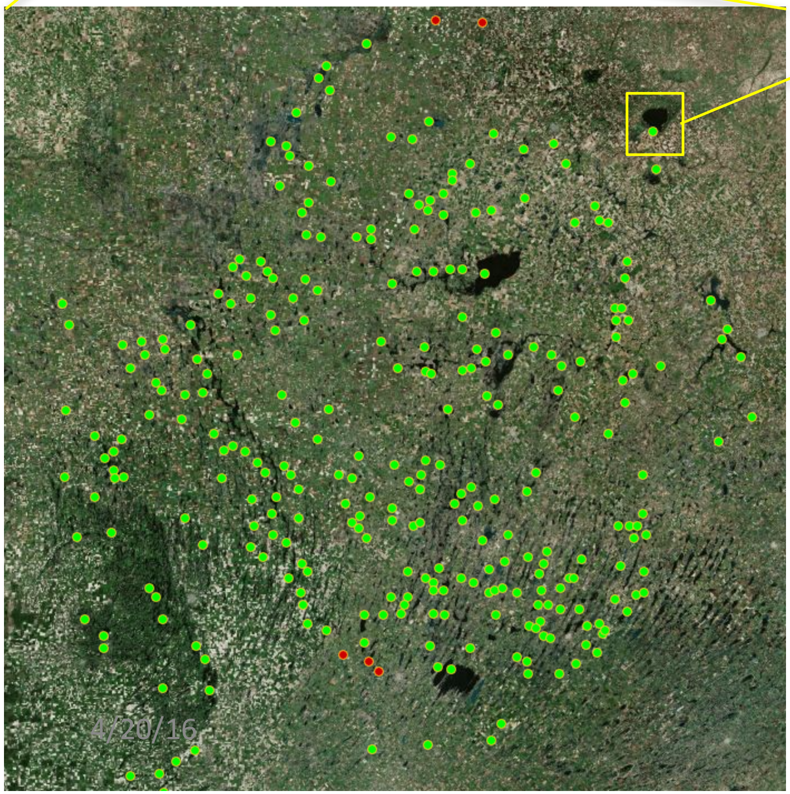


Example time series of a *Water Loss* region

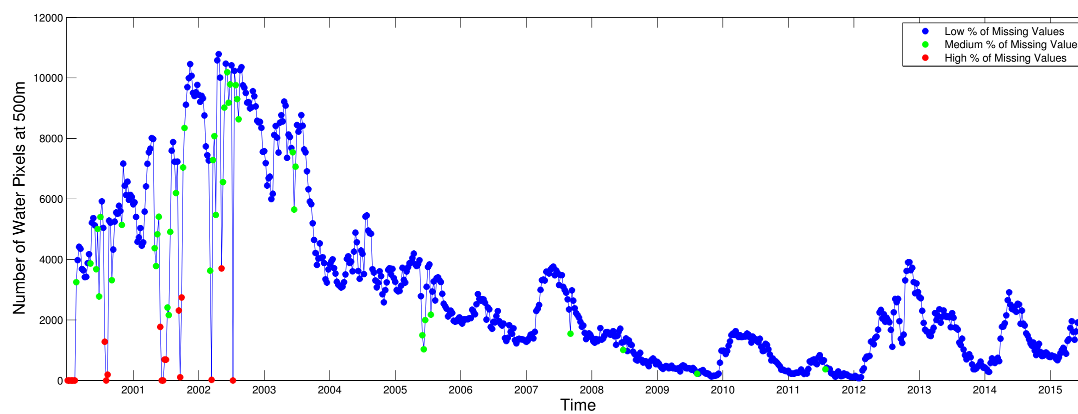
Examples of Change: Shrinking Water Bodies



(Green dots show regions changing from water to land in last 15 years)



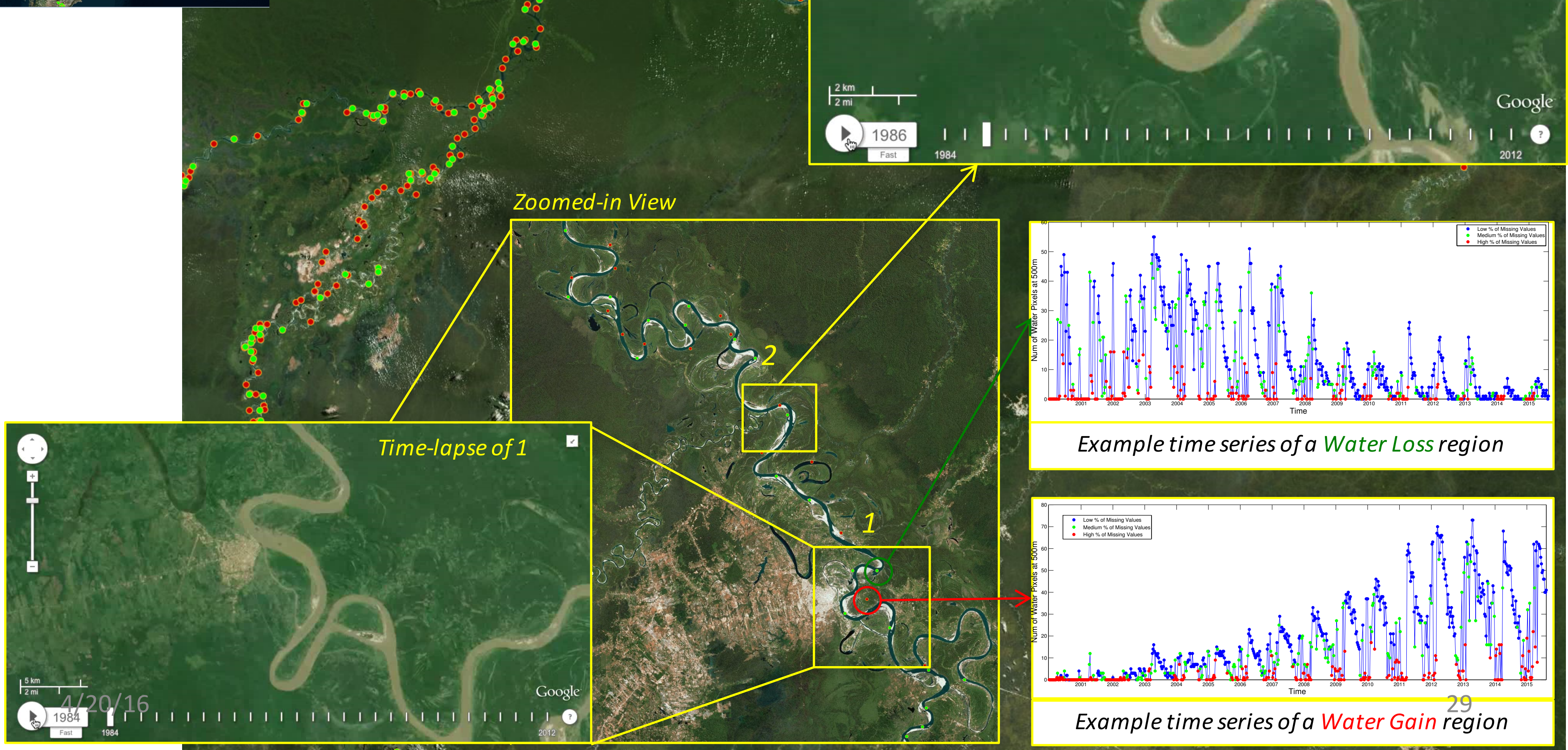
Annual Time-lapse of an example green dot



Aggregate dynamics of all green dots shown on left

Examples of Change: River Meandering

(Adjacent occurrence of *Water Gain (red)* and *Water Loss (green)* regions all along the river indicate the displacement of water from the green dots to the red dots)

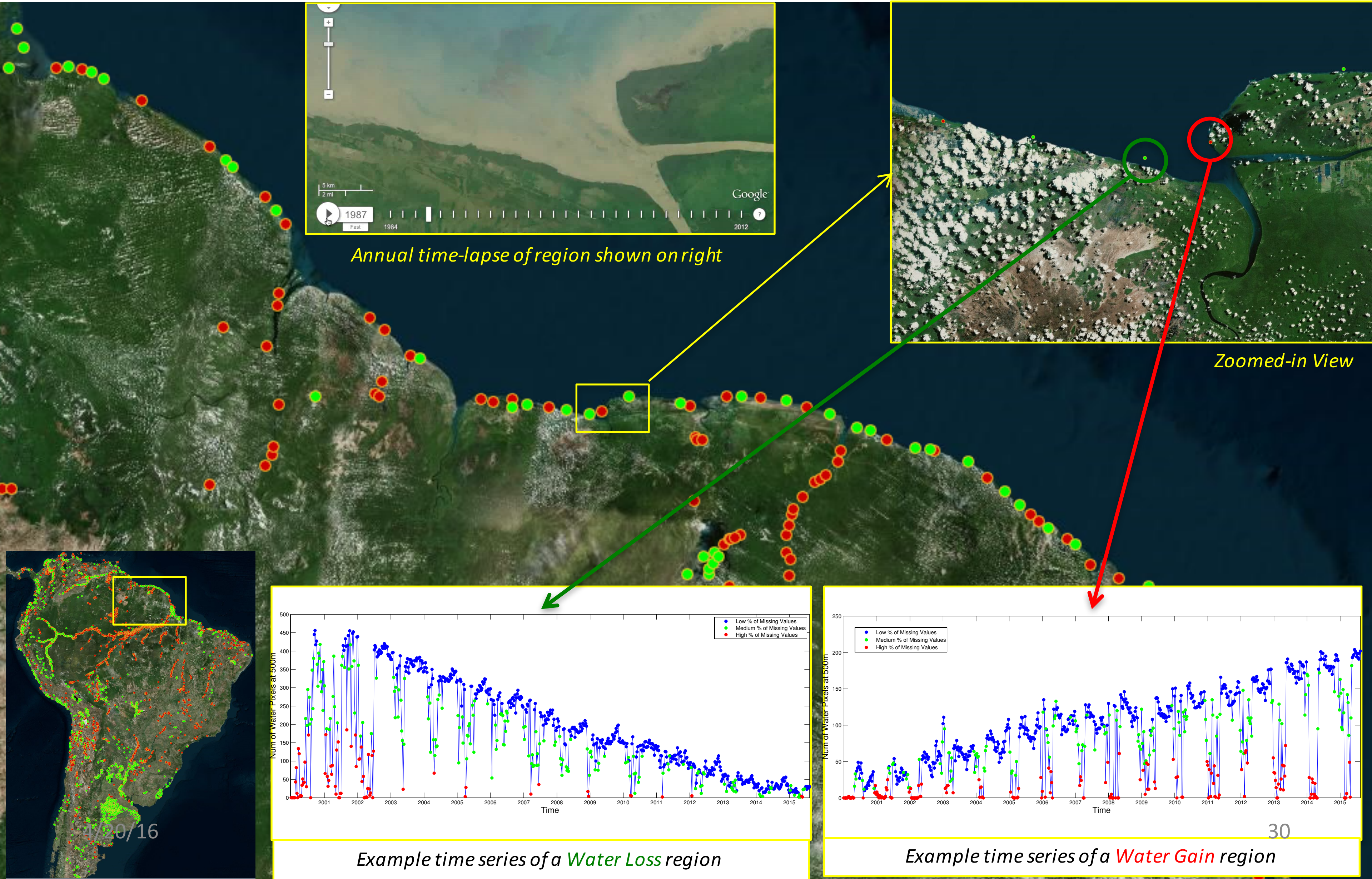


Example time series of a *Water Loss* region

Example time series of a *Water Gain* region

Examples of Change: Delta Erosion

(*Water Gain* and *Water Loss* regions appear on the coastline, due to displacement of sediments around river deltas)



Annual time-lapse of region shown on right

Zoomed-in View

Example time series of a Water Loss region

Example time series of a Water Gain region

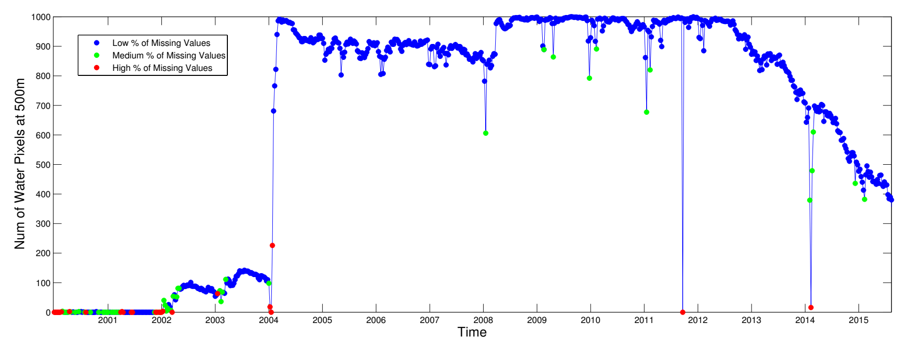
Examples of Change: Dam Constructions

Global Reservoir and Dam (GRanD) Database:

- A data curation initiative by Global Water System Project (GWSP)
- Finds dams constructed after 2001:
  - (**65** globally; **12** in Brazil)

UMN Approach:

- Finds (**458** globally; **134** in Brazil[1])



- Construction of a dam characterized by a sudden and persistent increase in surface area

[1]Prepared in collaboration with Juan Carlos, Planetary Skin Institute

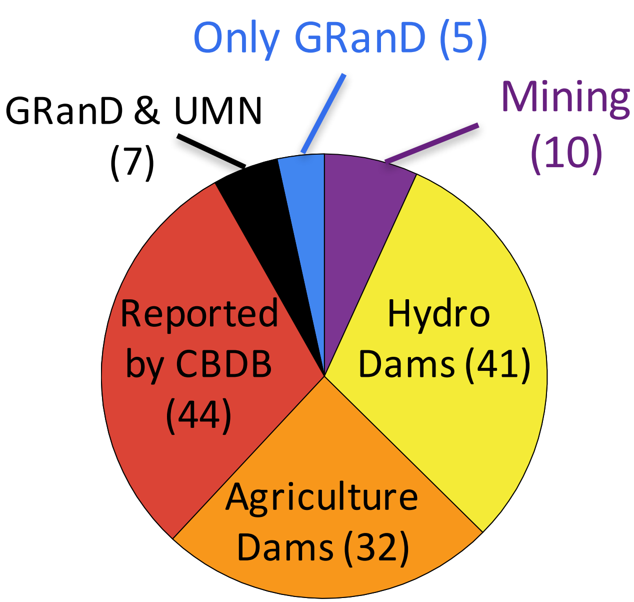
Examples of Change: Dam Constructions

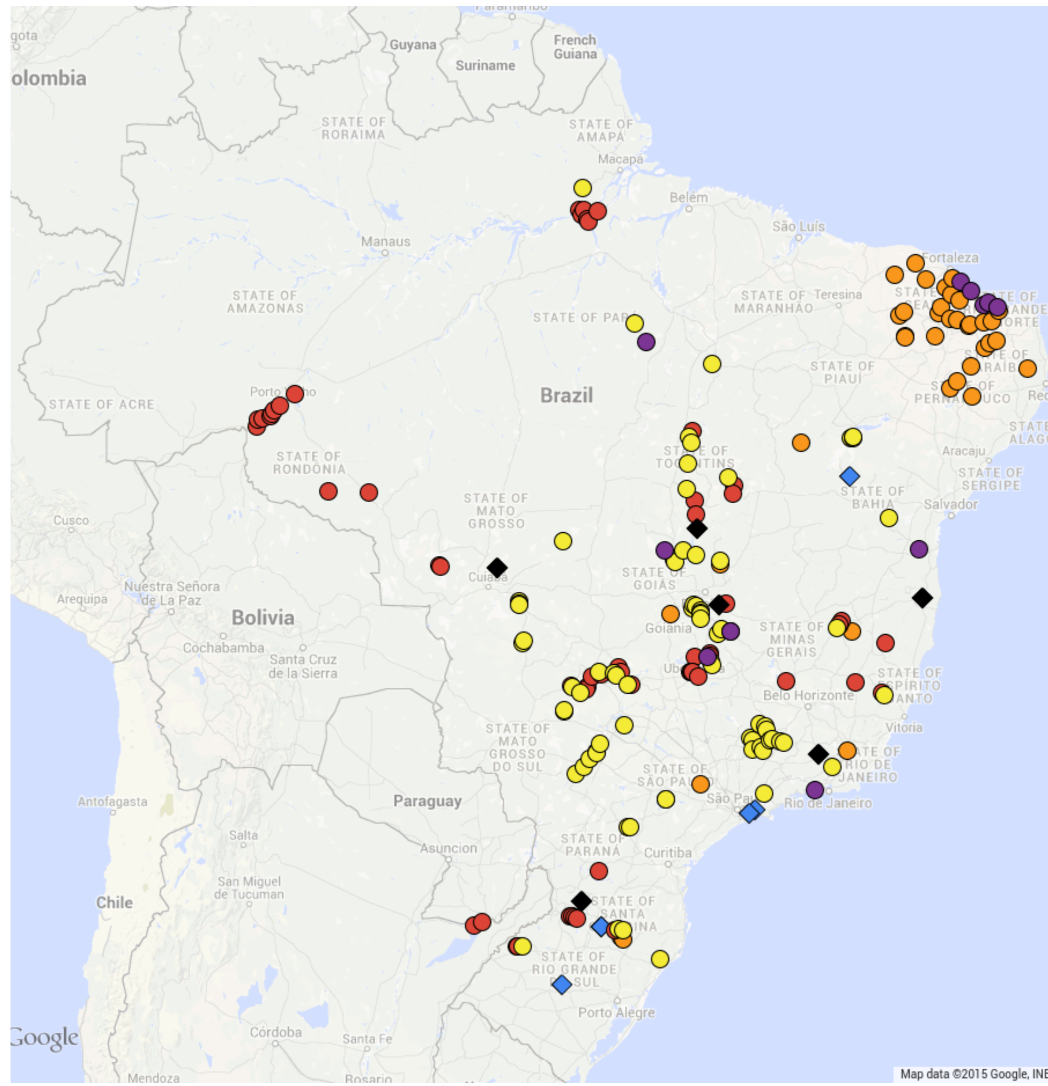
Global Reservoir and Dam (GRanD) Database:

- A data curation initiative by Global Water System Project (GWSP)
- Finds dams constructed after 2001:
  - (**65** globally; **12** in Brazil)

UMN Approach:

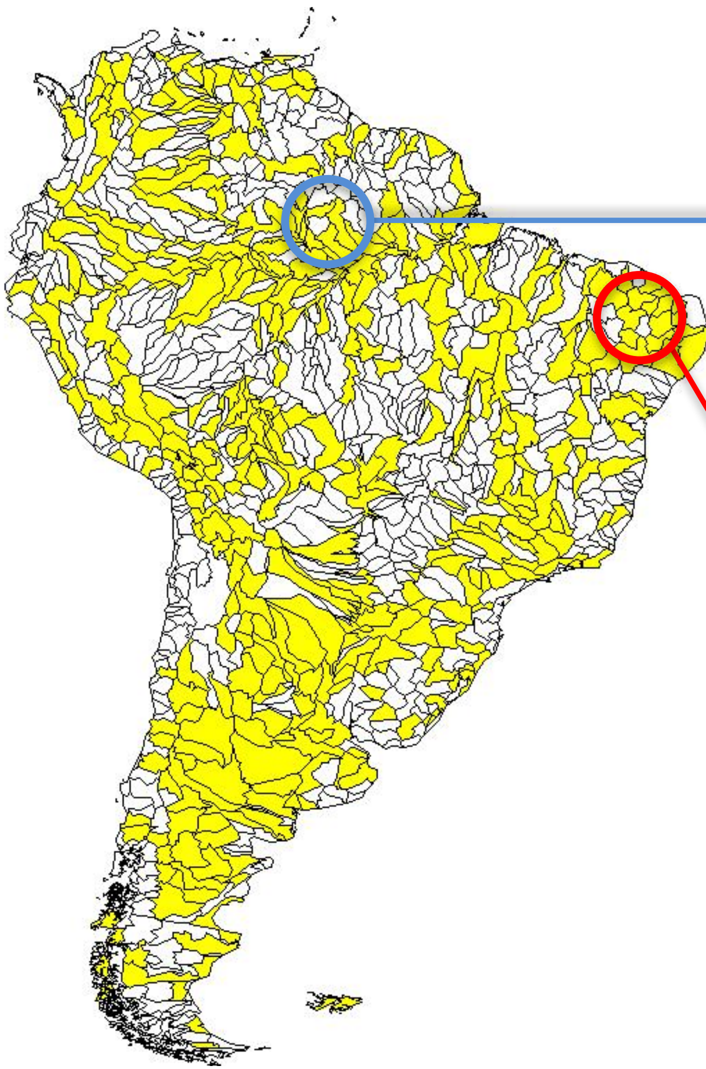
- Finds (**458** globally; **134** in Brazil[1])



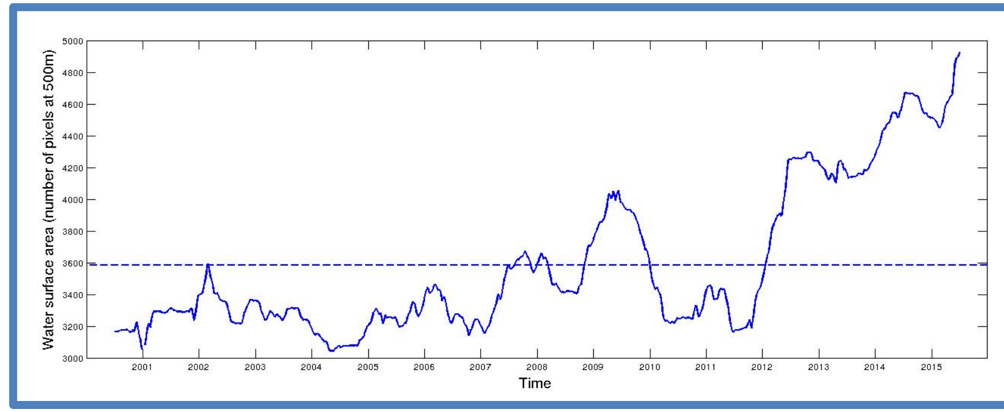


[1]Prepared in collaboration with Juan Carlos, Planetary Skin Institute

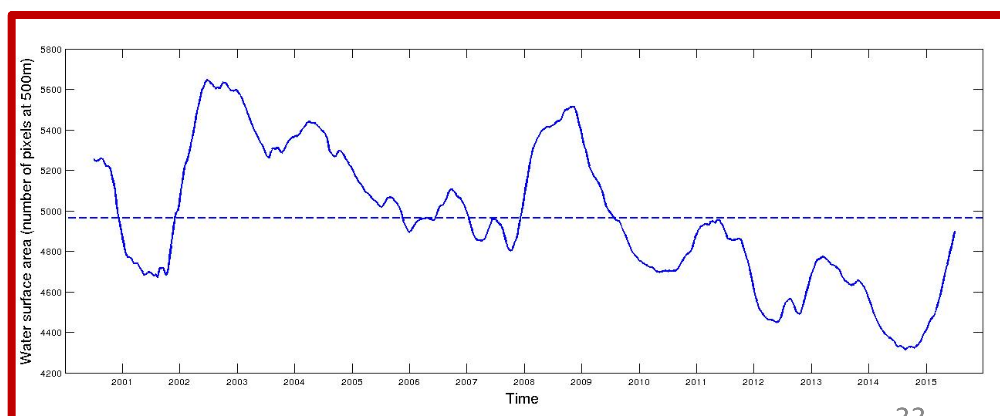
Aggregate Trends in Surface Water Dynamics



Surface Water Dynamics in Amazon



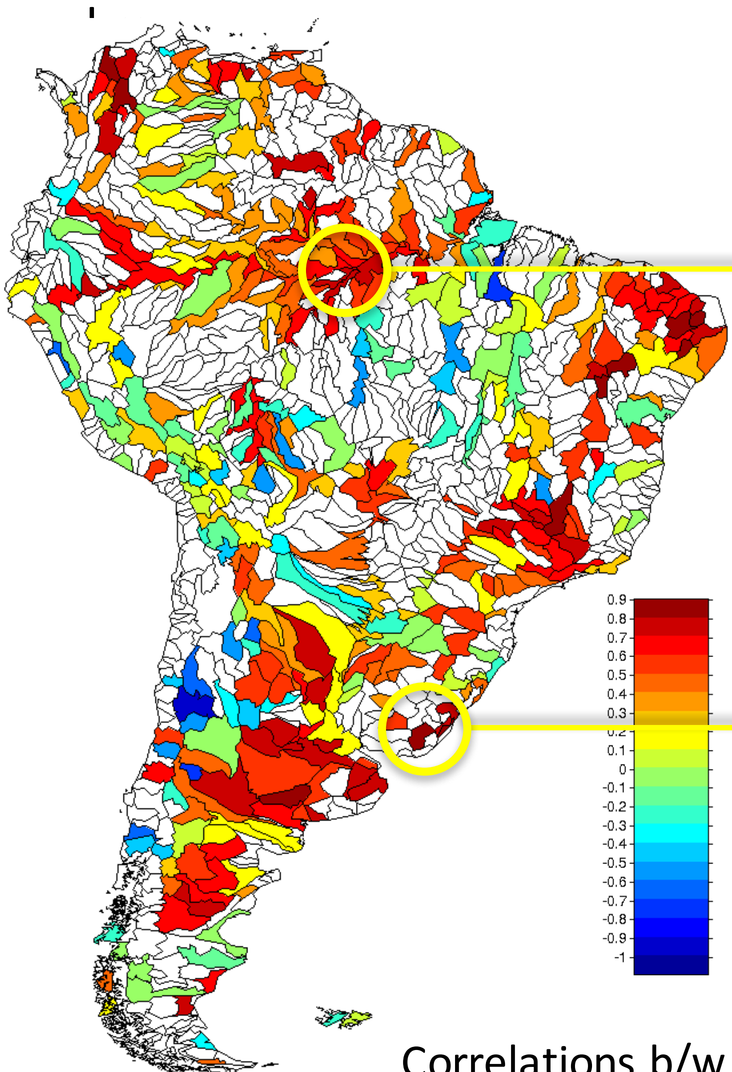
Surface Water Dynamics in NE Brazil

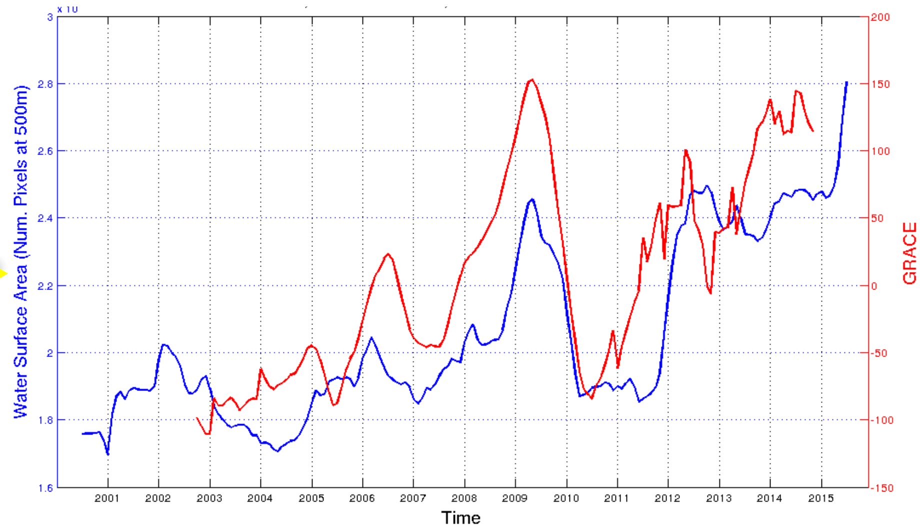


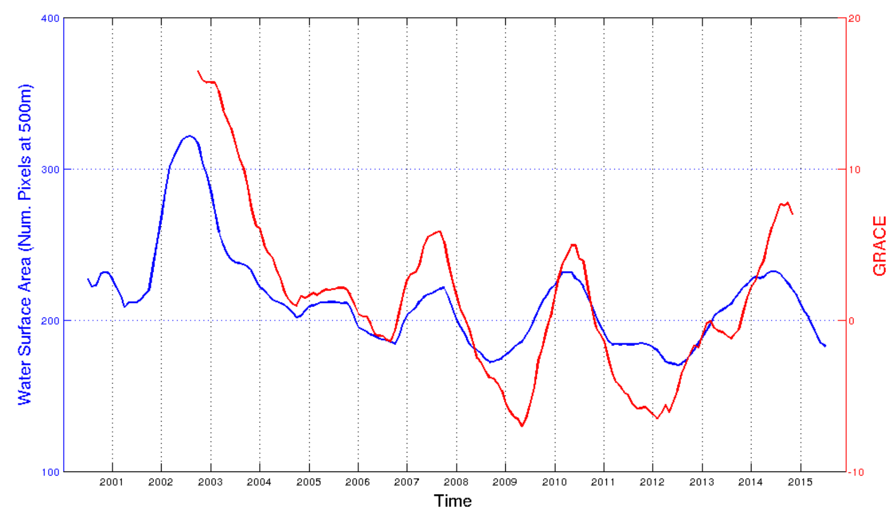
Correlations with GRACE

GRACE: Gravimetry Recovery and Climate Experiment

- Measures changes in total water mass (surface + groundwater) at ~100



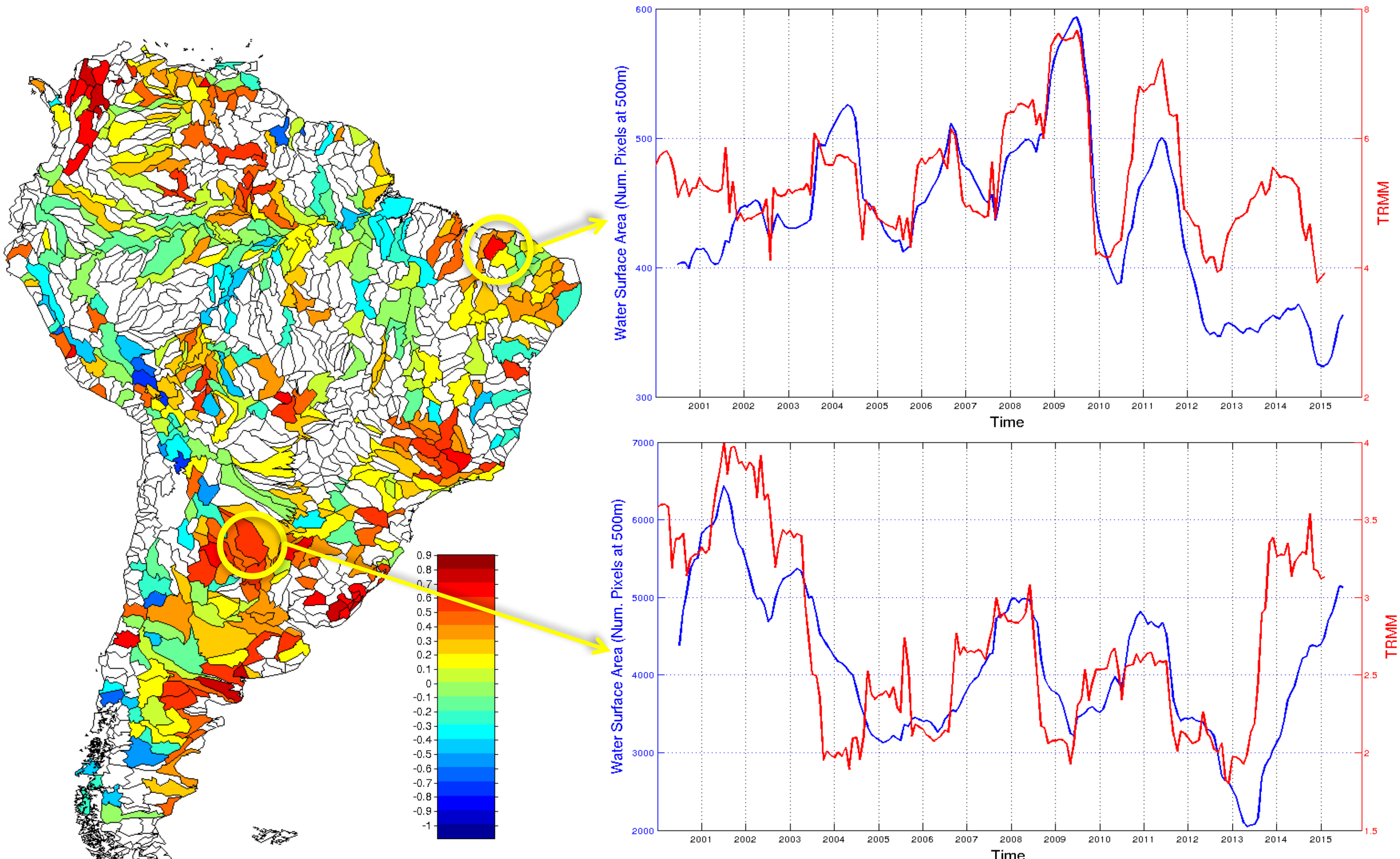




Correlations b/w surface water dynamics and GRACE measurements

Correlations with Precipitation

TRMM: Tropical Rainfall Measuring Mission (available at ~25 km)



Correlations b/w surface water dynamics and TRMM measurements

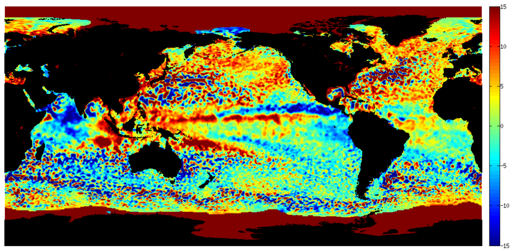
Potential Use Cases of a Water Monitoring System

- Quantifying water storage variations for all surface water bodies
  - Producing volume estimates of large lakes and reservoirs by integrating surface area extents with surface height measurements
- Building a comprehensive database of dams and reservoirs constructions at a global scale
- Studying the interactions between surface water dynamics and land cover changes, especially in the context of food-energy-water systems
- Mapping the dynamics of rivers and estimating their discharge at a global scale using fine-resolution Landsat data
- Integrating fine-scale information about surface water dynamics in hydrological models at regional to global scales

Five Year, $ 10m NSF Expeditions in Computing Project (1029711, PI: Vipin Kumar, U. Minnesota)

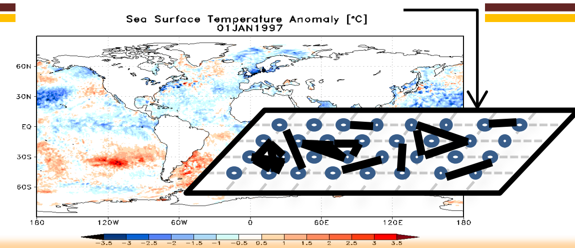
Understanding Climate Change: A Data-driven Approach

Research Highlights



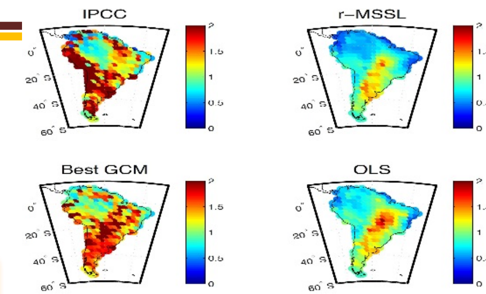
Pattern Mining: Monitoring Ocean Eddies

- Spatio-temporal pattern mining using novel multiple object tracking algorithms
- Created an open source data base of 20+ years of eddies and eddy tracks



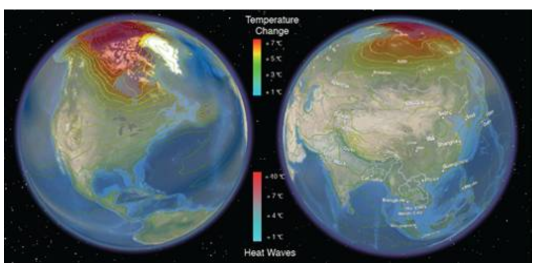
Highlights:

- Highly inter-displicinary
  - Computer science, hydrology, Earth sciences, statistics, civil engineering
- Dozens of publications (journals, conferences, and workshops) with authors from multiple disciplines
  - Papers in Nature and Nature Climate Change
- Public release of software & data products
- Advances in computer science driven by Earth science applications
- Advances in Earth sciences using computer science methods
- Development of physics-guided data mining paradigm



Sparse Predictive Modeling: Precipitation Downscaling

- Hierarchical sparse regression and multi-task learning with spatial smoothing
- Regional climate predictions from global observations

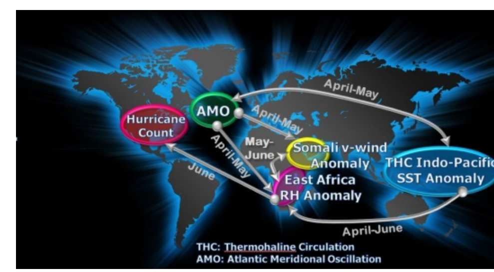


Extremes and Uncertainty: Heat waves, heavy rainfall

- Extreme value theory in space-time and dependence of extremes on covariates
- Spatiotemporal trends in extremes and physics-guided uncertainty quantification

Change Detection: Monitoring Ecosystem Distrubances

- Robust scoring techniques for identifying diverse changes in spatio-temporal data
- Created a comprehensive catalogue of global changes in surface water and vegetation, e.g. fires and deforestation.



Relationship mining: Seasonal hurricane activity

- Statistical method for automatic inference of modulating networks
- Discovery of key factors and mechanisms modulating hurricane variability

http://climatechange.cs.umn.edu/

Concluding Remarks

- Big data techniques hold great promise for increasing our understanding of the Earth's climate and environment.
- Domain theory can be used to guide the process of knowledge discovery in scientific data
  - "Theory-guided Data Science"
- Methods have applicability across diverse domains:
  - Ecosystem management
  - Epidemiology
  - Geospatial Intelligence
  - Neuroscience

Acknowledgements

Students

Graduate: Saurabh Aggrawal, Xi Chen, James Faghmous, Xiaowei Jia, Anuj Karpatne, Ankush Khandelwal, Varun Mithal, Guruprasad Nayak

Undergraduate: Reid Anderson, Eric Mccaleb, Robert Leuenberger, Yizheng Ding, Stryker Thompson, Mace Blank, Matthew Schultz, Shitong Song, Daniel Kim

NSF Expeditions Team Members

UMN:
Vipin Kumar, Arindam Banerjee, Shyam Boriah, Snigdhansu Chatterjee, Jonathan Foley, Joseph Knight, Stefan Liess, Shashi Shekhar, Peter Snyder, Michael Steinbach, Karsten Steinhaeuser

NCSU: Nagiza Samatova, Fredrick Semazzi
Northeastern: Auroop Ganguly
Northwestern: Alok Choudhary, Wei-keng Liao
North Carolina A&T: Abdollah Homaifar

External Collaborators
NASA Ames: Rama Nemani, Nikunj Oza, Christopher Potter
Institute on Environment, UMN: Kate Brauman, Kimberly Carlson, James Gerber, Jessica Hellmann
UCLA: Dennis Lettenmaier, Miriam Marlier
Global Water System Project: Bernhard Lehner
Cal State Monterey Bay: Stephen Klooster
Michigan State: Pang-Ning Tan
Planetary Skin Institute: Juan Carlos Castilla-Rubio

website: climatechange.cs.umn.edu

Publications

- A. Karpatne and S. Liess. "A Guide to Earth Science Data: Summary and Research Challenges." *Computing in Science & Engineering* 17.6 (2015): 14-18.
- A. Karpatne, Z. Jiang, R. R. Vatsavai, S. Shekhar, and V. Kumar. "Monitoring Land Cover Changes using Remote Sensing Data: A Machine Learning Perspective," *IEEE Geoscience and Remote Sensing Magazine*, 2016.
- J. Faghmous and V. Kumar. "A big data guide to understanding climate change: The case for theory-guided data science." *Big data* 2.3 (2014): 155-163.
- J. Faghmous, A. Banerjee, S. Shekhar, M. Steinbach, V. Kumar, A. Ganguly, N. Samatova. "Theory-guided data science for climate change." *IEEE Computer* 11 (2014): 74-78.
- A. Karpatne, A. Khaldelwal, X. Chen, V. Mithal, J.H. Faghmous, and V. Kumar. "Global monitoring of inland water dynamics: State-of-the-art, challenges, and opportunities." *In K. Morik, J. Lässig, and K. Kersting, Eds. Computational Sustainability, Springer.* 2016.
- A. Karpatne and V. Kumar. "Adaptive heterogeneous ensemble learning using the context of test instances", *International Conference on Data Mining (ICDM)*, 2015
- A. Khandelwal, V. Mithal, and V. Kumar. "Post-classification label refinement using implicit ordering constraints among data instances." *International Conference on Data Mining (ICDM)*, 2015.
- X. Chen, J. Faghmous, A. Khandelwal, and V. Kumar. "Clustering dynamic spatio-temporal patterns in the presence of noise and missing data." *International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.
- A. Karpatne, A. Khandelwal, and V. Kumar. "Ensemble learning methods for binary classification with multi-modality within the classes." *SIAM International Conference on Data Mining (SDM)*, 2015
- A. Karpatne, A. Khandelwal, S. Boriah, and V. Kumar. "Predictive Learning in the Presence of Heterogeneity and Limited Training Data." *SIAM International Conference on Data Mining (SDM)*, 2014.
- V. Mithal. "Learning with uncertain and incomplete data." PhD Dissertation, 2016.