BIG DATA IN CLIMATE AND EARTH SCIENCES: OPPORTUNITIES AND CHALLENGES

Xiaowei Jia

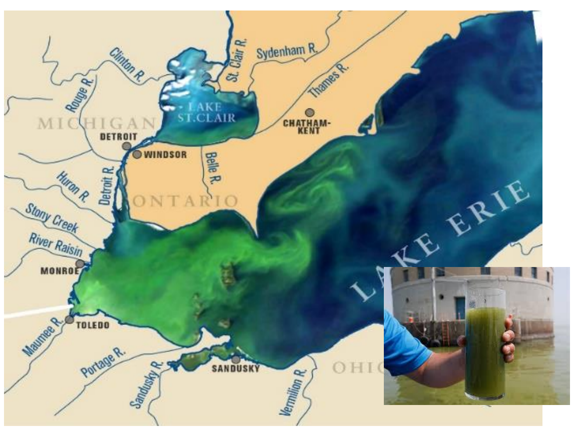
University of Pittsburgh

xiaowei@pitt.edu

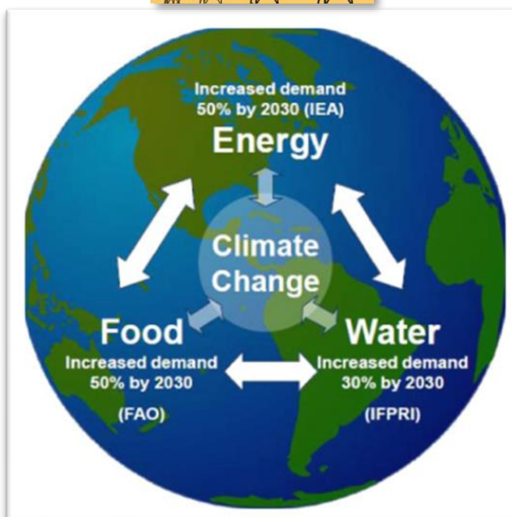
Environmental Grand Challenges of the 21st Century

Major Climate Report Describes a Strong Risk of Crisis as Early as 2040,
New York Times, Oct. 7, 2018

Floods due to Hurricane Harvey



Harmful Algal Bloom in Lake Erie

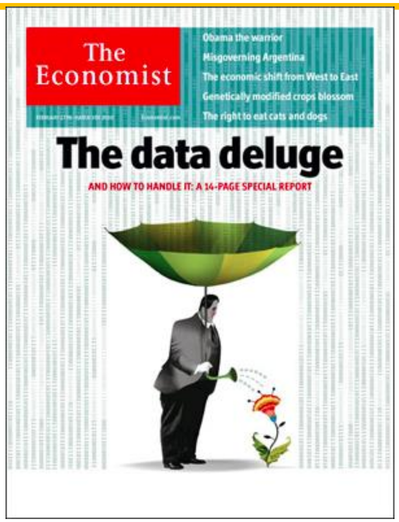


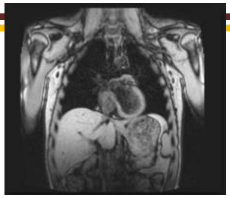
How to Feed the World Without Killing the Planet?
Cool Green Science by Nature, July 7, 2017

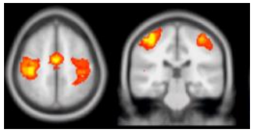
Oil Palm Plantations in Indonesia

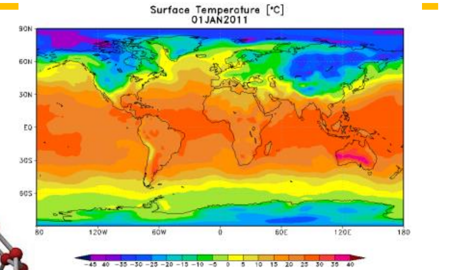
Shrinking Salt Lake

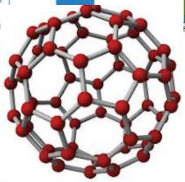
Golden Age of Data Science

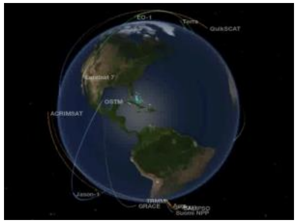


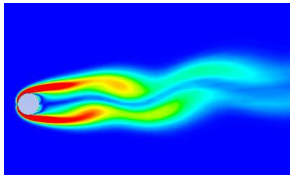












- Hugely successful in commercial applications:

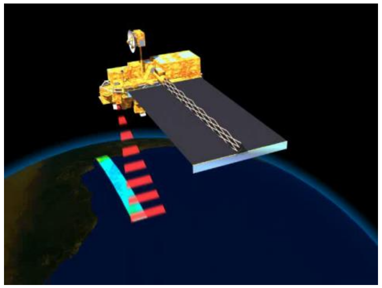
DeepMind

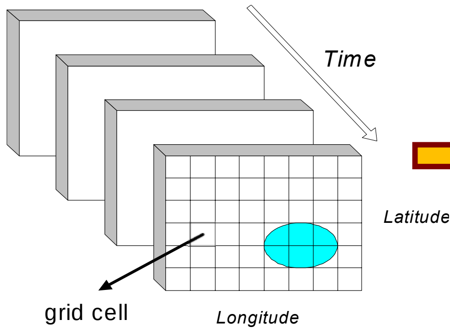
Google Ads

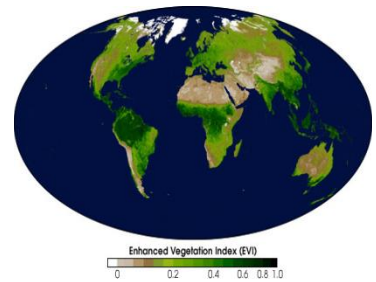
facebook

NETFLIX

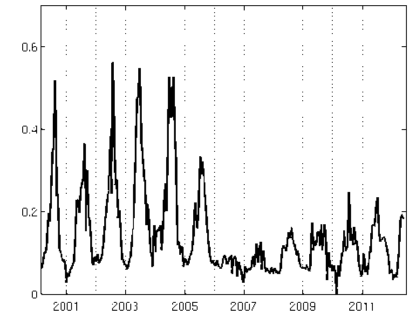
Big Data in Earth System Monitoring







A **vegetation index** measures th 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. Mapping of plantation dynamics in tropical forests:

- ❑ Recurrent Neural Networks to model space and time

3. Global mapping of inland surface water dynamics

- ❑ Heterogeneous Ensemble Learning
- ❑ Physics-guided Labeling
- ❑ Information Transfer across Space and Time

Lake Oroville in 2011 and 2014

Case Study 1: Global Forest Fires Mapping

Monitoring fires is important for climate change impact

A record number of more than 150 countries signed 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.

8000 sq.Km scene in SE Asia(2005)

Burned area mapping

Predicts whether a given pixel is burned or not?

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:

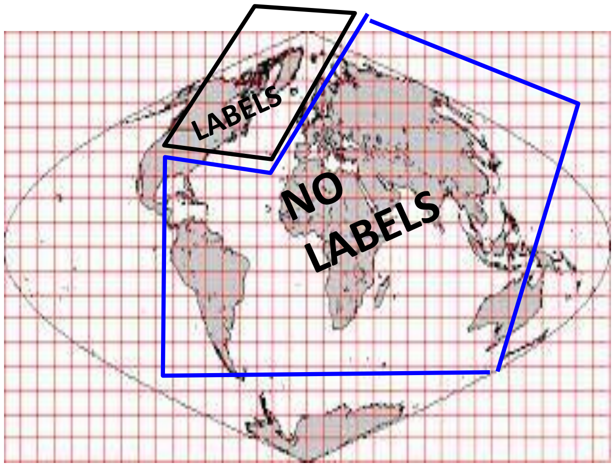
(1) Complete absence of target labels for supervision

(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)

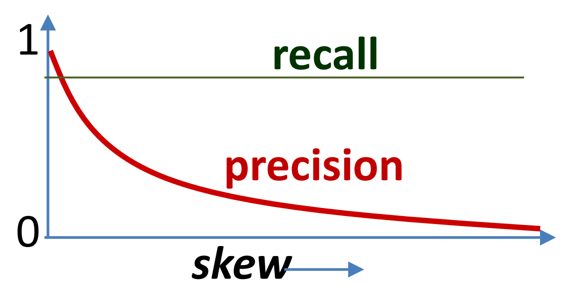
(2) Highly imbalanced classes

For eg. **California State**

Year 2008 had very high fire activity

2,296 sq. km. of forests burned
out of a total
73,702 sq. km. forested area

True Positive Rate = 0.9
False Positive Rate = 0.01



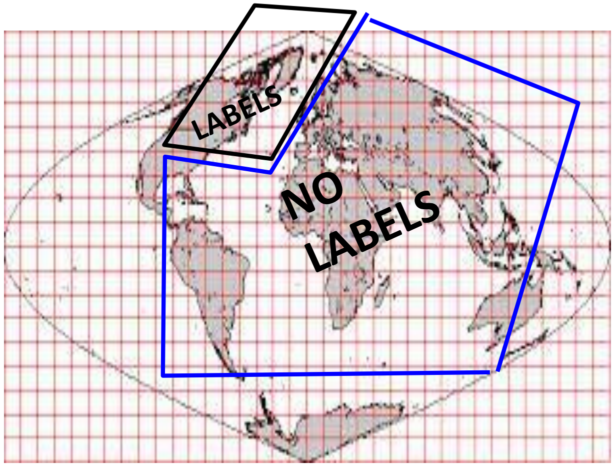
Predictive Modeling for Global Monitoring of Forest Fires

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?

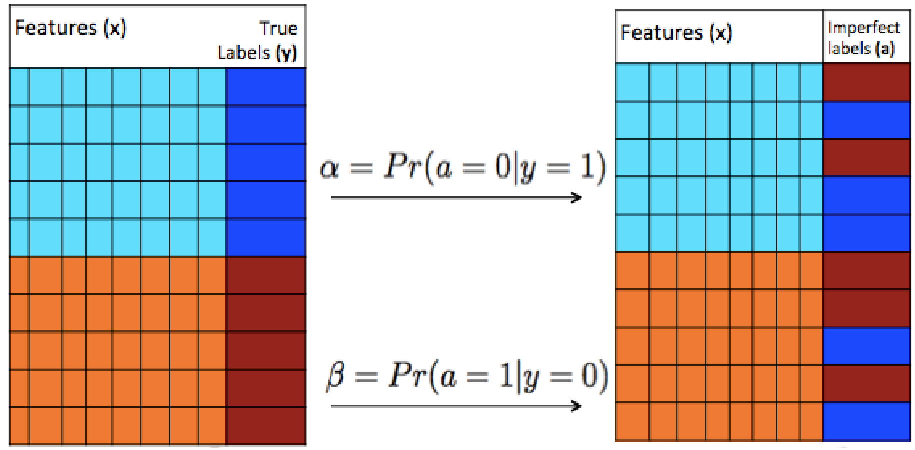


Global availability of labeled samples for burned area classification

Predictive Modeling for a Rare Target Class using Imperfect Labels

What are imperfect labels ?

- Noisy/perturbed true labels
- Inexpensive to obtain
  - Raw feature
  - Heuristics (given by expert)
- Available for all test instances



Examples of imperfect labels

Application	Target label	Imperfect label
Burned Area	Fire/No Fire	Thermal anomaly
Urban settlement	Urban/No urban	Night time light
Recommending items to a new user	Interested/ No interested	Friends interest

Learning with imperfect labels

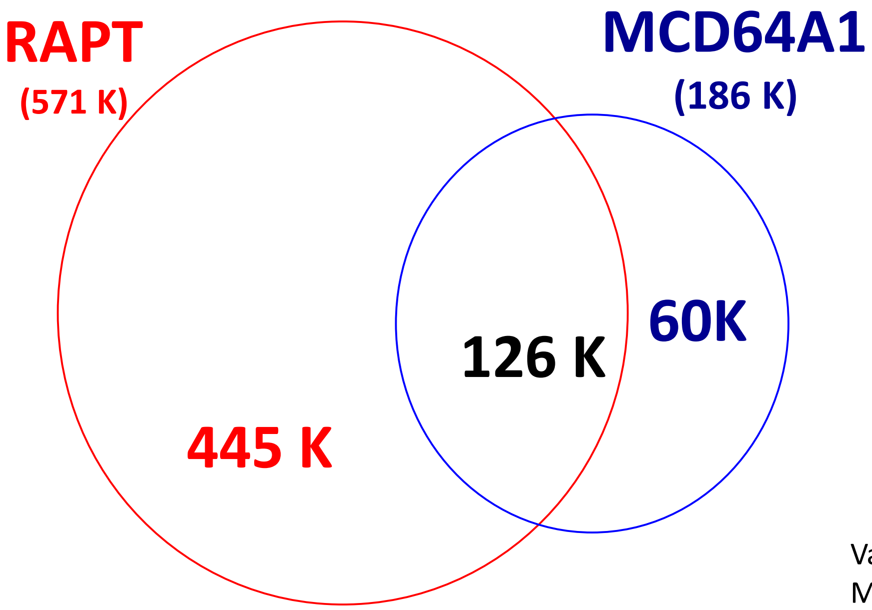
- **Supervised Learning**
  - **Expert-annotated Labels**
    - **Sufficient training samples**
      - *SVM*
      - *Decision tree*
      - *Logistic regression*
      - *Deep Learning*
    - **Inadequate training samples**
      - *Semi-supervised*
      - *Active Learning*
      - *Multi-view*
      - *Multi-task*
  - **Imperfect Labels**
    - **Multiple annotators**
      - *Learning with crowds*
      - *Raykar et al.*
    - **Single annotator**
      - **Partial Supervision**
        - *Positive Unlabeled learning*
        - *Bing Liu et al.*
        - *Elkan et al.*
      - **Imperfect Supervision**
        - **Balanced**
          - *Natrajan et al.*
          - *Menon et al.*
        - **Rare class**

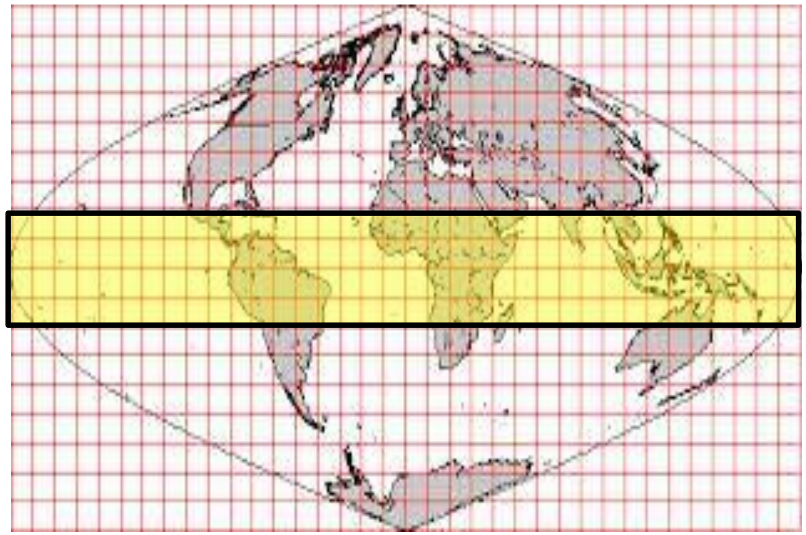
Global Monitoring of Fires in Tropical Forests

Fires in tropical forests during 2001-2014

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

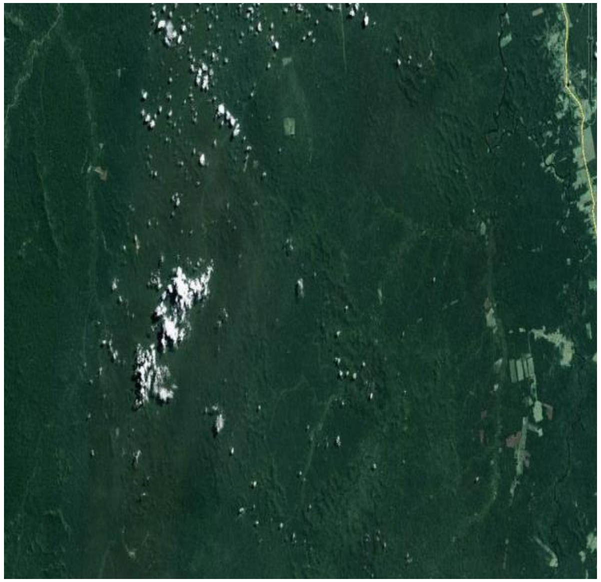
three times the area reported by state-of-art NASA product: MCD64A1





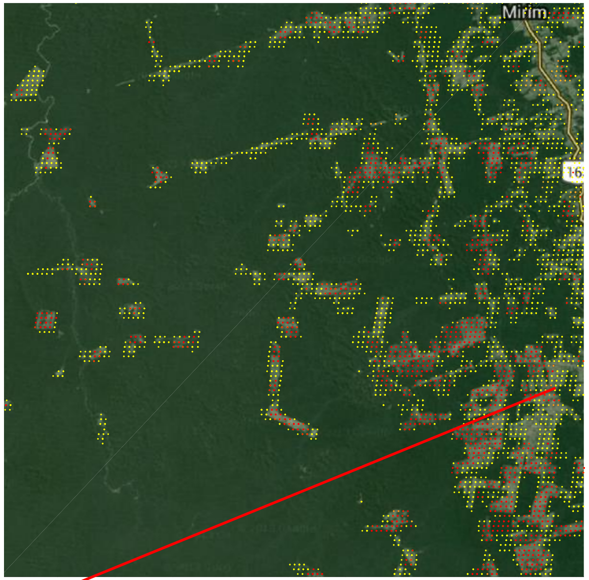
Varun Mithal et. al, "Mapping Burned Areas in Tropical Forests Using a Novel Machine Learning Framework." *Remote Sensing* 10, no. 1 (2018): 69.

Deforestation via Burning 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

Monitoring Global Change: Case Studies

1. Global mapping of forest fires:
    - RAPT: Rare Class Prediction in Absence of Ground Truth

2. Mapping of plantation dynamics in tropical forests:
    - Recurrent Neural Networks to model space and time

3. Global mapping of inland surface water dynamics
    - Heterogeneous Ensemble Learning
    - Physics-guided Labeling
    - Information Transfer across Space and Time

Case Study 2: Mapping of Plantation Dynamics

SCIENTIFIC AMERICAN SUSTAINABILITY

Palm Oil Set to Grow Indonesia's Climate Changing Emissions

Draining peatlands and replacing forests with palm oil plantations may cause Indonesian pollution to soar, despite pledges

By Nathanael Massey, ClimateWire on October 10, 2012

nature International weekly journal of science

Home | News & Comment | Research | Careers & Jobs | Current Issue | Archive | Audio & Video

NATURE | NEWS

Fibre production drives deforestation in Indonesia

Study debunks belief that palm-oil plantations are main culprit.

Natasha Gilbert 21 July 2014

SCIENTIFIC AMERICAN SUSTAINABILITY December 1, 2012

Stop Burning Rain Forests for Palm Oil

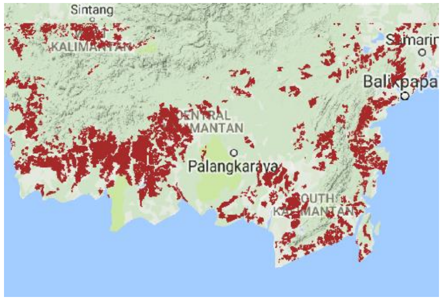
The world's growing appetite for cheap palm oil is destroying rain forests and amplifying climate change

Interplay between food, energy and water:

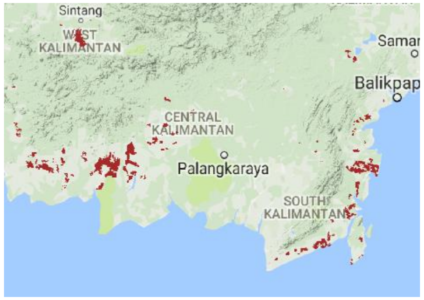
Production of edible oils and biofuels.
High carbon emissions.
Degradation of water quality.

State-of-the-art and Challenges

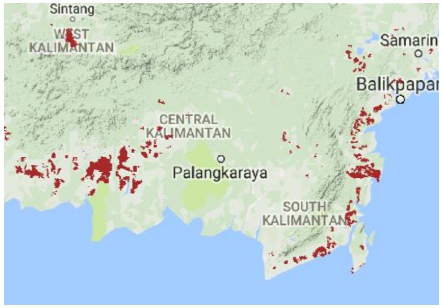
- **Tree Plantation (TP):** This data set is created by Transparent World, with the support of Global Forest Watch. Plantations are manually annotated on 2014. *TP has high recall and low precision.*
- **Roundtable on Sustainable Palm Oil (RSPO):** This dataset is available across Indonesia in 2000, 2005, and 2009. In addition, the study digitized all the locations into 19 land cover types in these eras. *RSPO has high precision and low recall.*



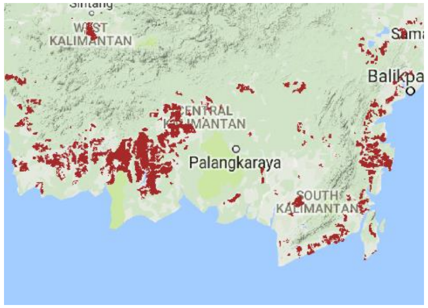
TP, 2014



RSPO, 2001



RSPO, 2005



RSPO, 2009

➢ Challenges

- Imperfect annotators
  - *Tree Plantation (TP):* high recall and low precision.
  - *Roundtable on Sustainable Palm Oil (RSPO):* high precision and low recall.
- Data heterogeneity
  - Land cover heterogeneity
    - ❑ differentiate plantation from a variety of land covers, e.g. forest, are highly confused with plantations.
  - Spatial heterogeneity
  - Temporal heterogeneity
    - ❑ Seasonal variation - e.g., a crop land after harvest looks very similar to a barren land.
    - ❑ Yearly variation – the spectral features of a land cover change across years.
- Noisy and high-dimensional feature space

Our Contribution

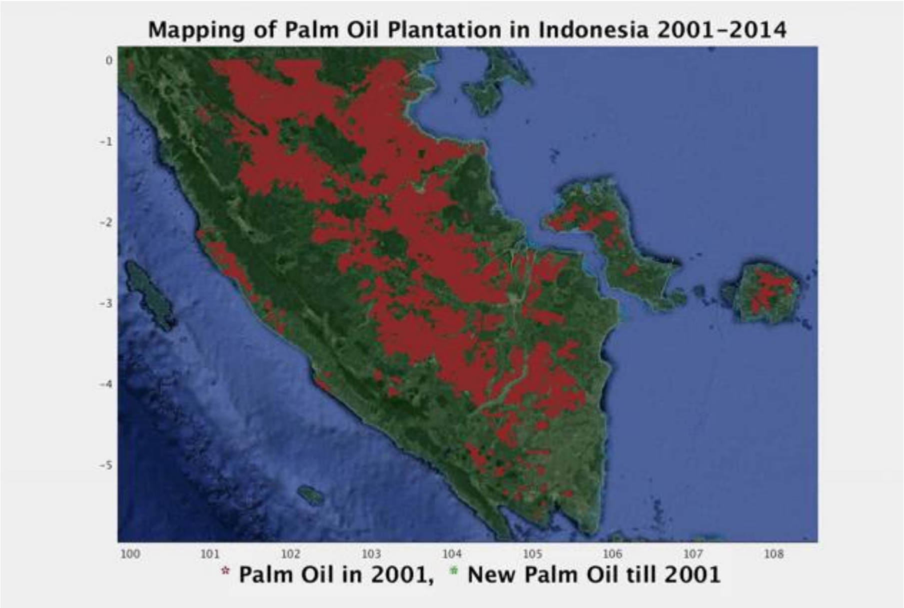
- Learning from multiple imperfect annotators
  (Jia et al. BigData 2016)
- Each annotator has different expertise level on different plantation types.
- We recursively update the expertise of each annotator and estimate true labels.

- Handling temporal heterogeneity in prediction
  (Jia et al. SDM 2017)
- We model temporal and spatial dependencies across years in an LSTM model.
- We propose an incremental learning strategy to update the LSTM model.

- Aggregating classes, collecting samples and validating results
  (Jia et al. Technical Report, 2017)
- We aggregate similar classes according to domain expertise.
- For each aggregated class, we sample equal amount of samples from each sub-class across multiple years.
- We validate the generated plantation maps by comparing random sampled locations to high-resolution images.

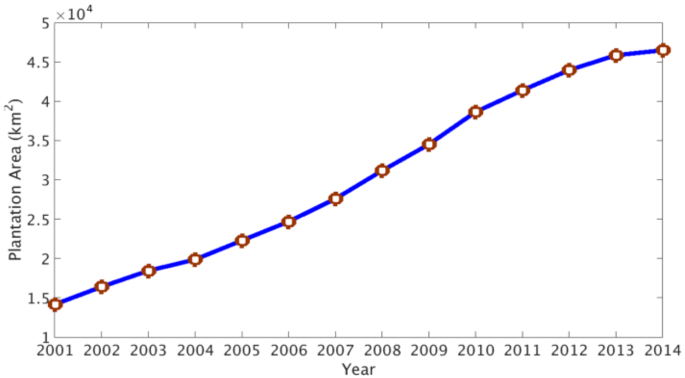
- 1. Jia, X., Khandelwal, A., Gerber, J., Carlson, K., West, P., and Kumar, V. Learning Large-scale Plantation Mapping from Imperfect Annotators. In IEEE Big Data (Big Data), 2016.
- 2. Jia, X., Khandelwal, A., Nayak, G., Gerber, J., Carlson, K., West, P., and Kumar, V. Predict Land Covers with Transition Modeling and Incremental Learning. In SDM, 2017.
- 3. Jia, X., Khandelwal, A., Gerber, J., Carlson, K., Samberg, L., West, P., and Kumar, V. Automated Plantation Mapping in Southeast Asia Using Remote Sensing Data. In Department of Computer Science and Engineering-Technical Reports.

Annual Plantation Maps

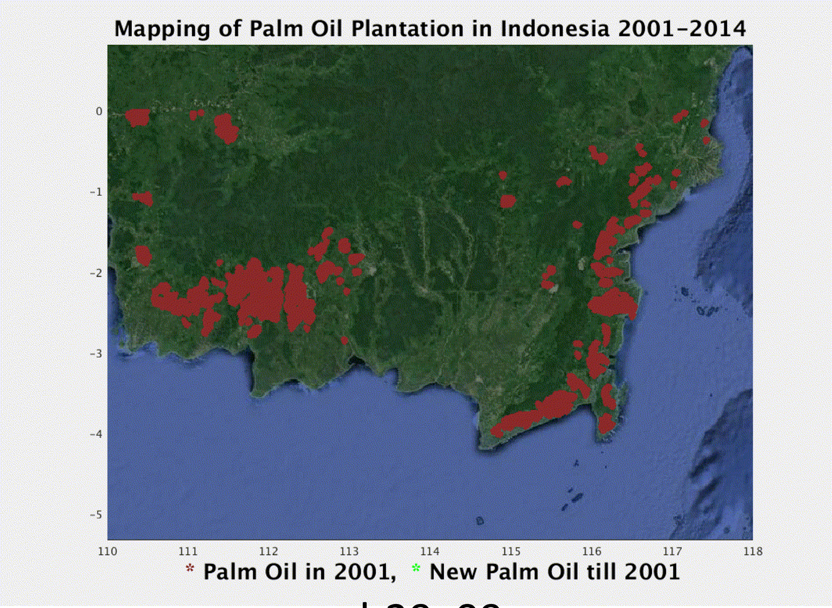


h28v09

h29v08

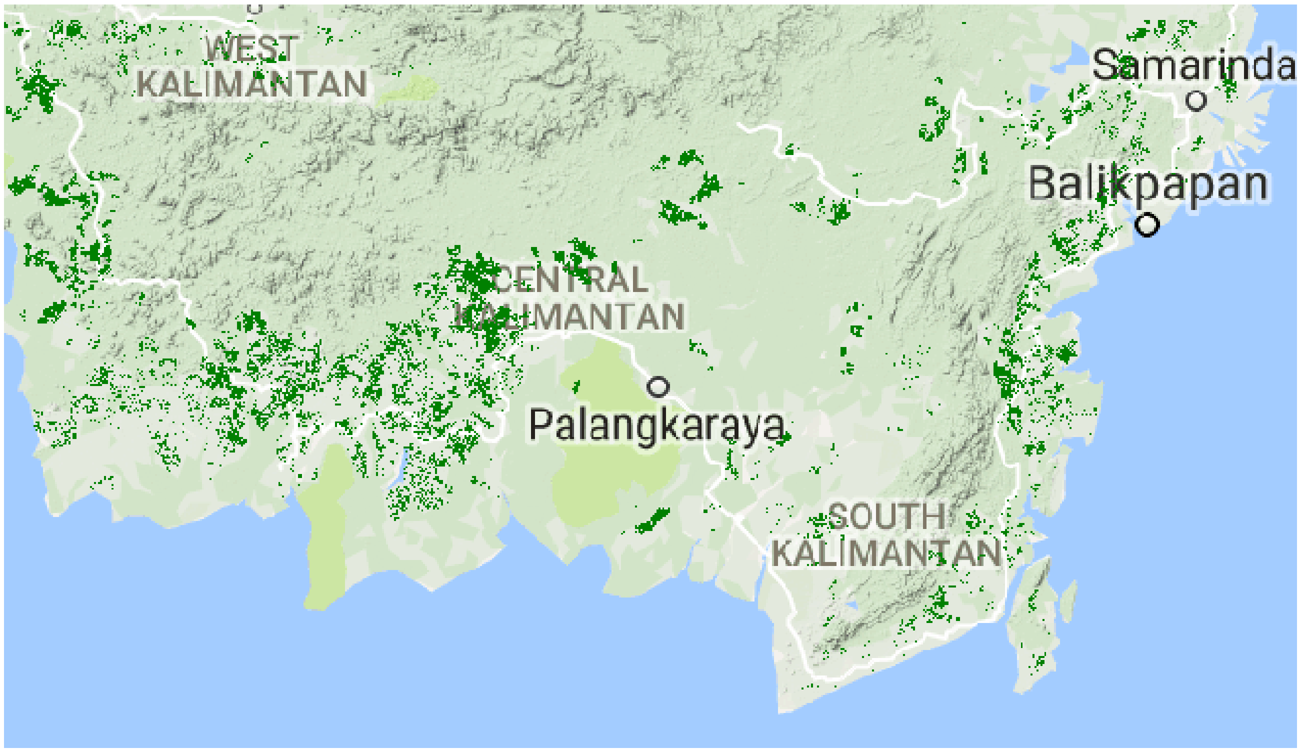


- Annual growth rate ≈ 9.57%



h29v09

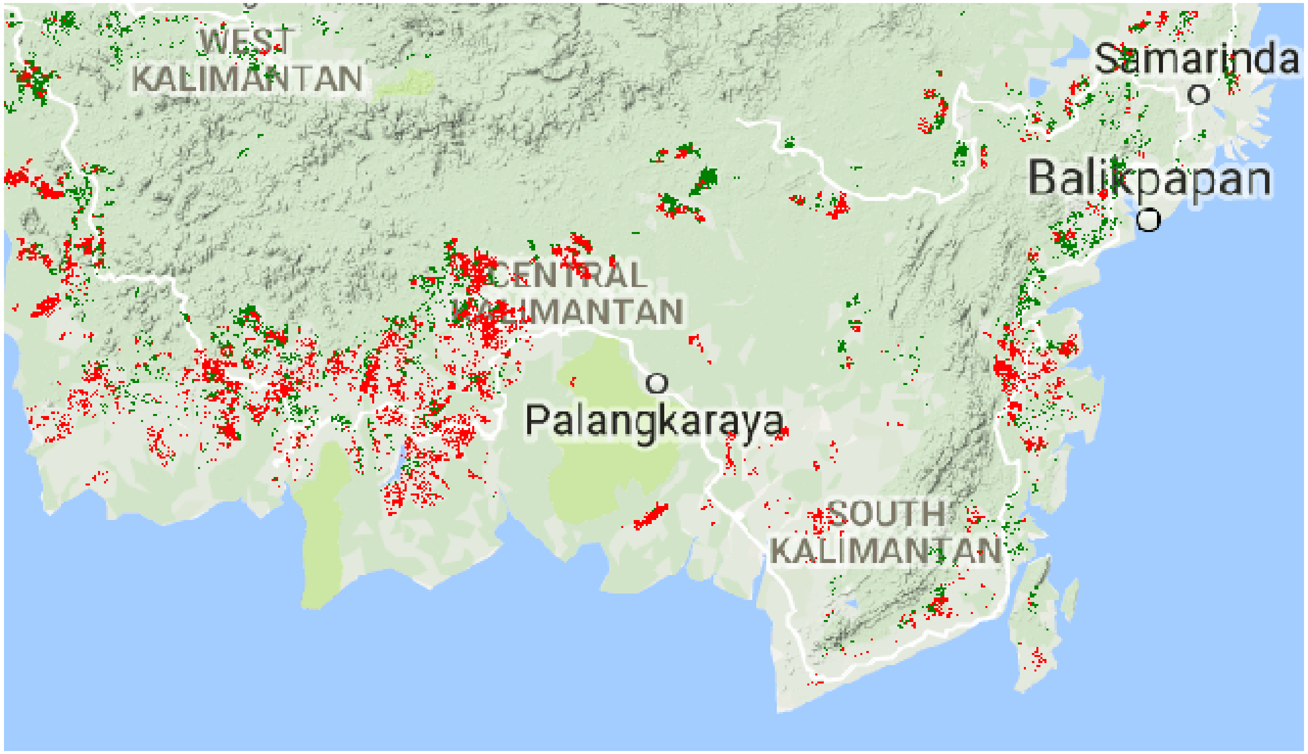
Interaction between Fires and Palm Oil Plantation



All plantations

This and all following figures show only confident forest pixels.

Interaction between Fires and Palm Oil Plantation

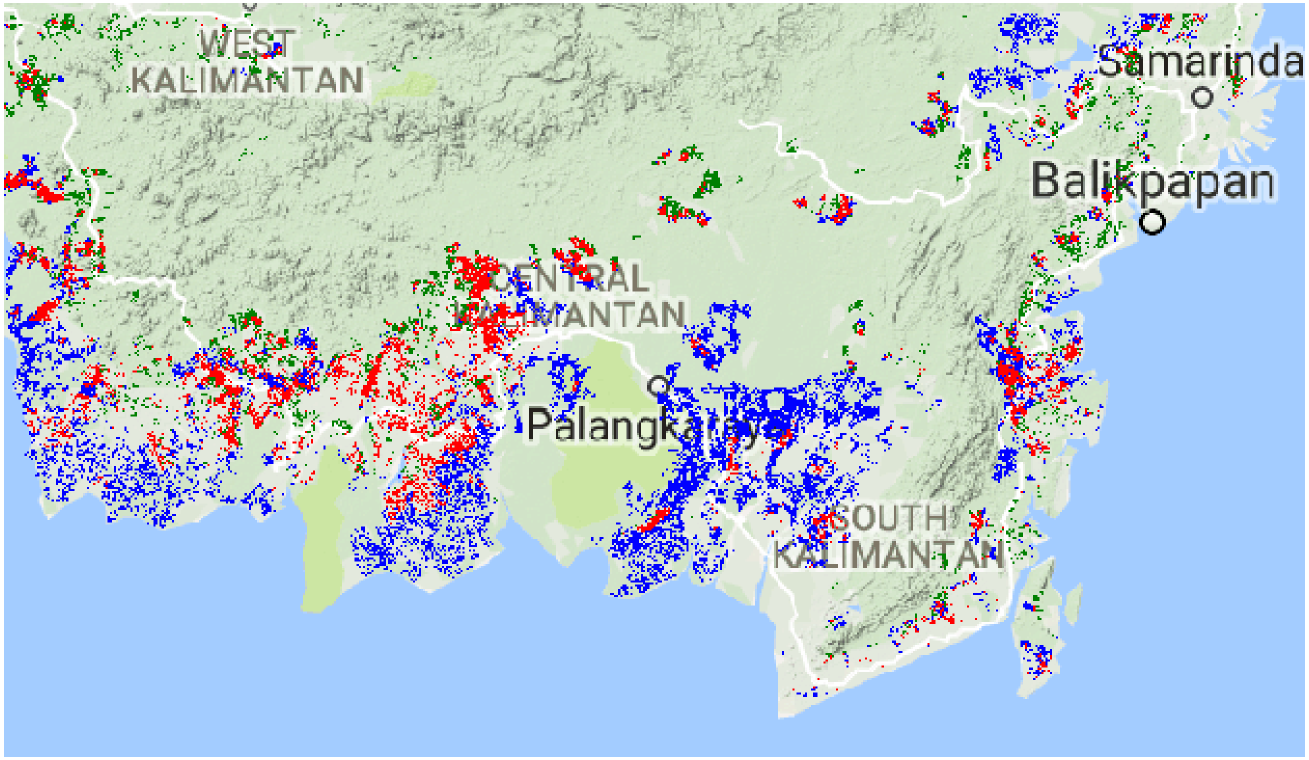


Plantations with no burn scars

Plantations with burn scar during 2001-2014

This and all following figures show only confident forest pixels.

Interaction between Fires and Palm Oil Plantation



Plantations with no burn scars

Plantations with burn scar during 2001-2014

Burned pixels around plantations (same burn date as nearby red pixels)

This and all following figures show only confident forest pixels.

Monitoring Global Change: Case Studies

1. Global mapping of forest fires:
   - RAPT: Rare Class Prediction in Absence of Ground Truth

2. Mapping of plantation dynamics in tropical forests:
   - Recurrent Neural Networks to model space and time

3. Global mapping of inland surface water dynamics
   - Heterogeneous Ensemble Learning (SDM 2015, ICDM 2015)
   - Physics-guided Labeling (ICDM 2015, RSE 2017)
   - Information Transfer across Space and Time (Khandelwal PhD Thesis)

Lake Oroville in 2011 and 2014

Importance of Monitoring Global Surface Water Dynamics

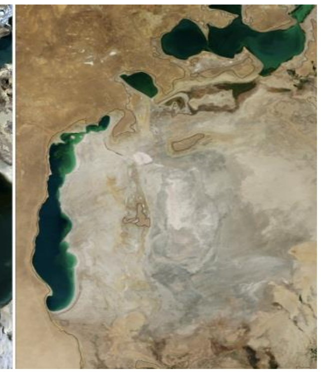
Impact of Climate Change

Cedo Caka Lake in Tibet, 1984

Cedo Caka Lake in Tibet, 2011

Impact of Human Actions

Aral Sea in 1989



Aral Sea in 2014

Early Warning Systems

Great Flood of Mississippi River, 1993

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?

Importance of Monitoring Global Surface Water Dynamics

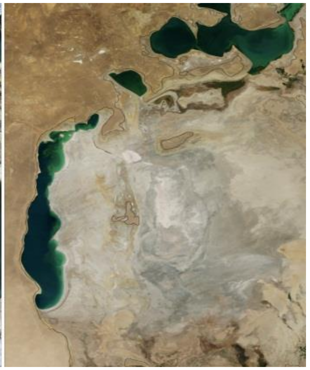
Impact of Climate Change

Cedo Caka Lake
in Tibet, 1984

Cedo Caka Lake
in Tibet, 2011

Impact of Human Actions

Aral Sea in 1989

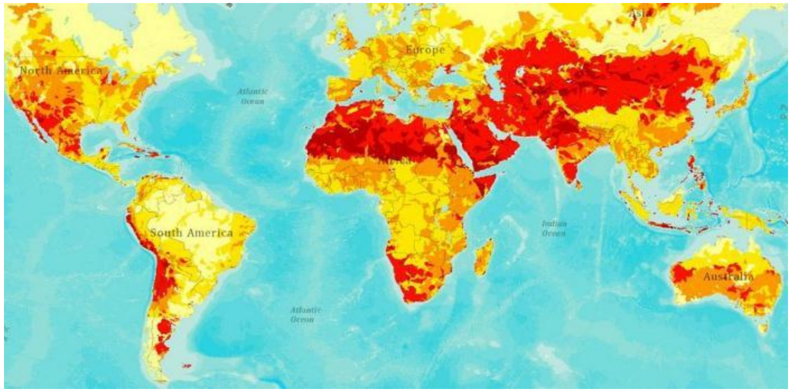


Aral Sea in 2014

Early Warning Systems

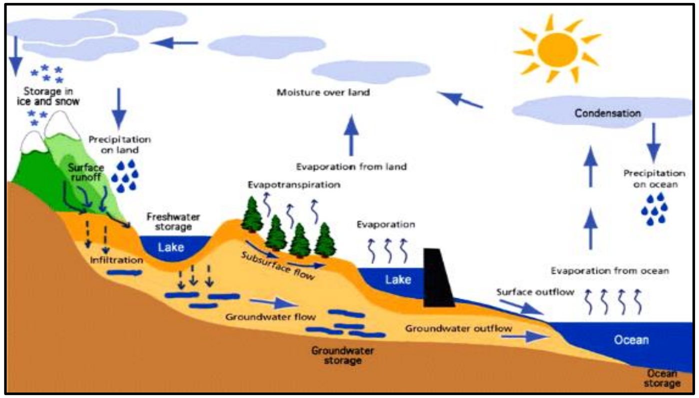
Great Flood of Mississippi
River, 1993

Quantifying water stocks and flow



Global projections of water risks (red)

Integrating with hydrological models



Importance of Monitoring Global Surface Water Dynamics

Impact of Climate Change

Cedo Caka Lake in Tibet, 1984

Impact of Human Actions

Early Warning Systems

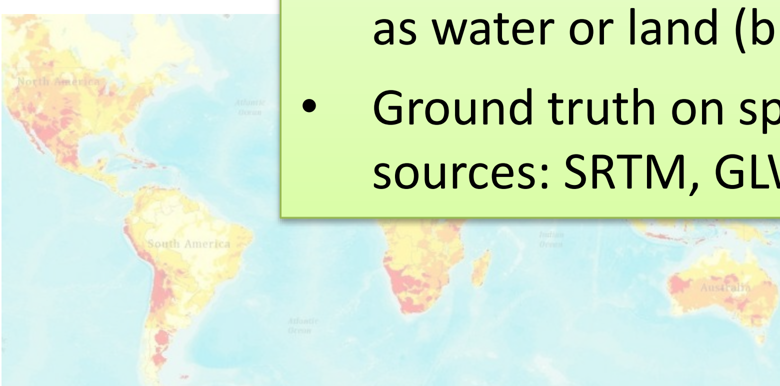
f Mississippi
1993

Remote Sensing Data

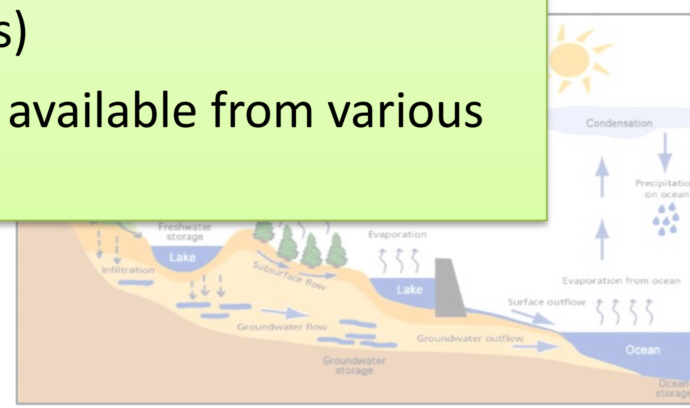
- Multi-spectral data
  - MODIS (at 500m, from 2000)
  - Landsat (at 30m, from 1970s)
- Can be used to classify every location at a given time as water or land (binary classes)
- Ground truth on specific dates available from various sources: SRTM, GLWD

Quantifyir

models

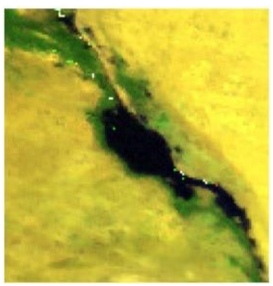


Global projections of water risks (red)



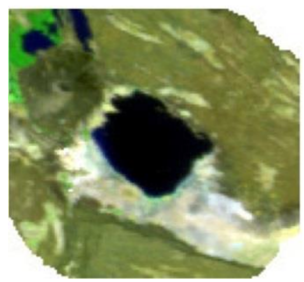
Challenges for Traditional Big Data Methods in Monitoring Water

- **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

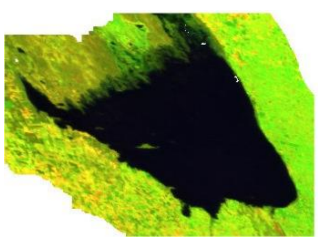


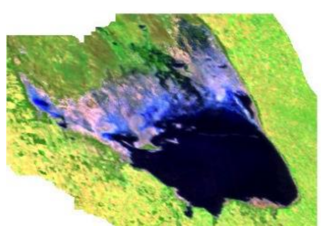
Great Bitter Lake, Egypt

Lake Tana, Ethiopia



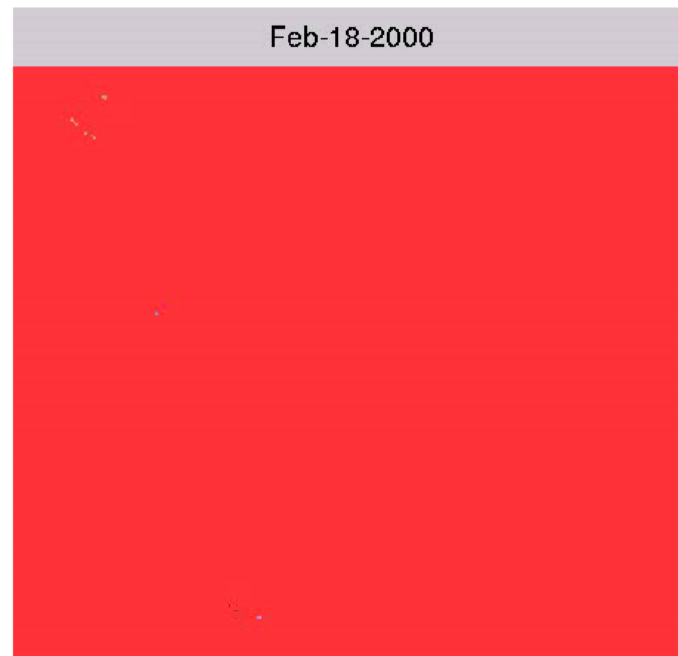
Lake Abbe, Africa





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

- **Challenge 2: Data Quality**
  - Clouds, shadows, atmospheric disturbances
    - Missing data – no labels
    - Incorrect labels



Poyang Lake, China

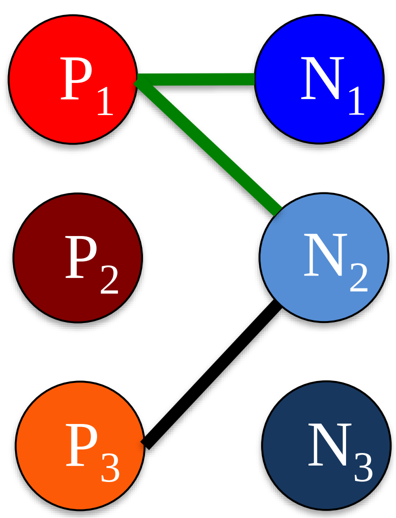
(Pink color shows missing data)

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

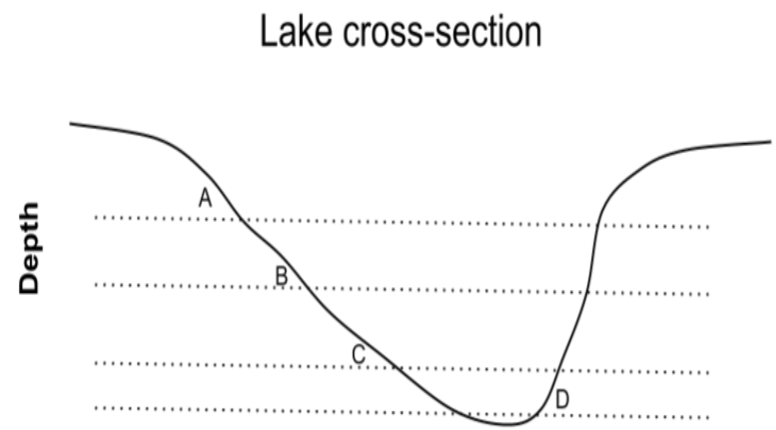
Positive Modes (Water) Negative Modes (Land)



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

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

ORBIT **(Ordering Based Information Transfer)** uses elevation information to constrain physically-consistent labels



Elevation A > B > C > D

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

A Global Surface Water Monitoring System

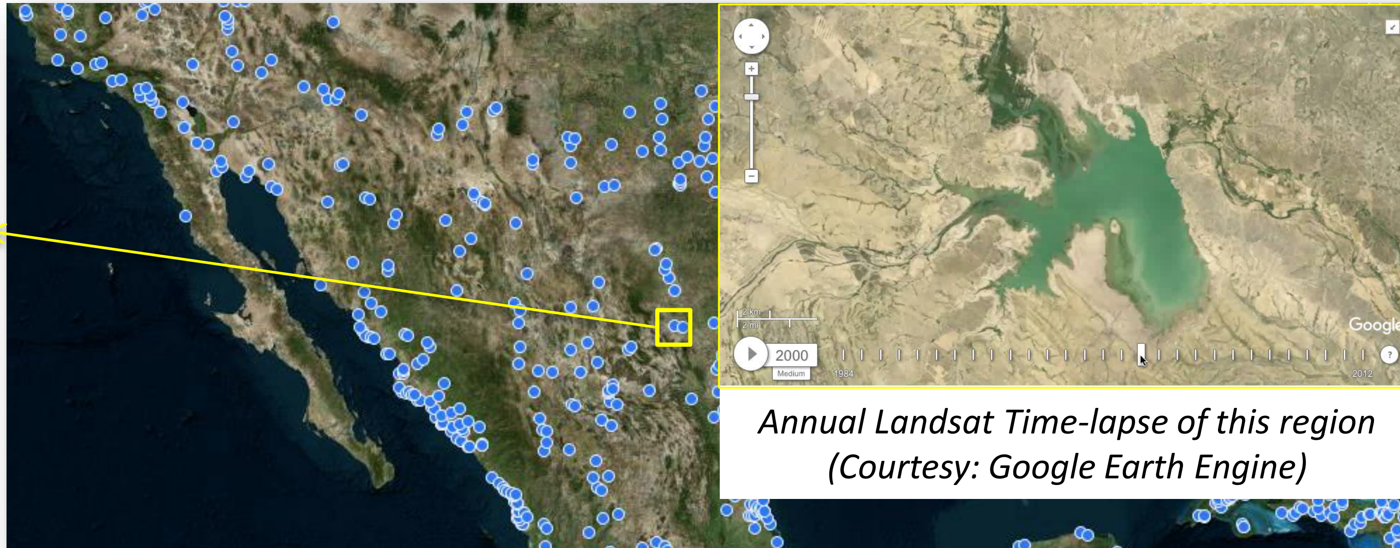
Maps the dynamics of all major surface water bodies (surface area > 2.5 km^2) shown as *blue dots*

<u>Key features</u>

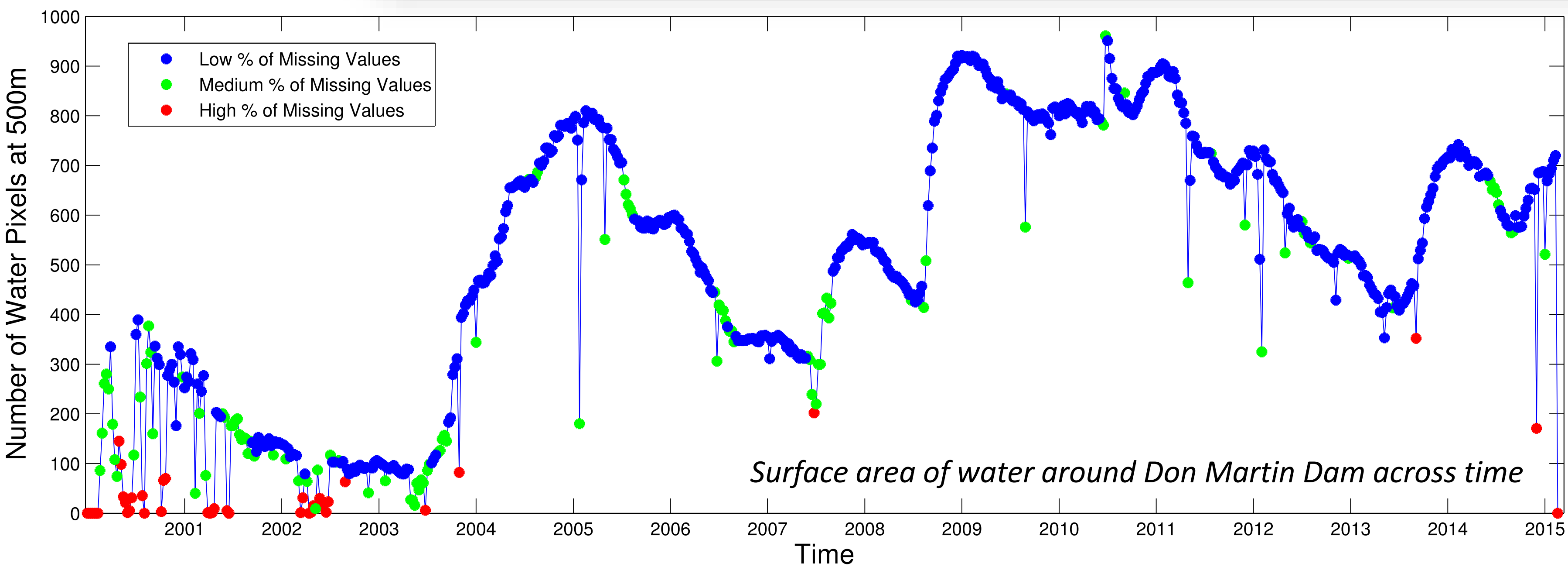
- *Identification of waterbodies*
  - *500m MODIS-based surface extent maps at 8 day interval*
  - *In progress: Daily time series of change at 30m resolution*
- *Local-regional change analysis*
- *Detects melting of glaciers/lakes*
- *Maps changes in river morphology*
- *Identifies reservoir constructions*
- *Finds relationships b/w surface water and precipitation/groundwater*

Showing Surface Water Dynamics

Don Martin Dam, Mexico



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

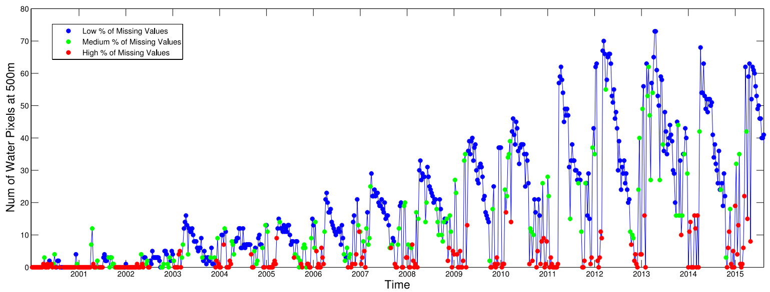


Surface area of water around Don Martin Dam across time

Regions of Change in South America

Red Dots *(Water Gain):*

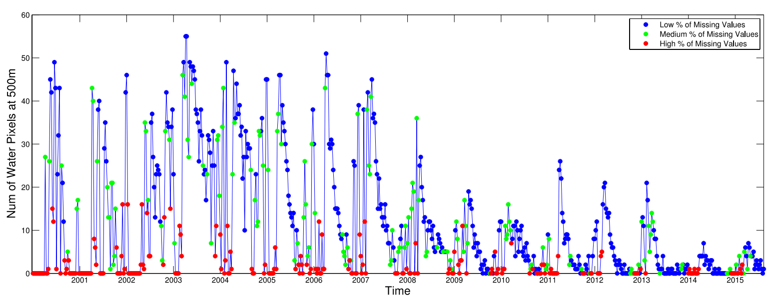
Region of size > 2.5 km^2 that have changed from land to water in the last 15 years



Example time series of a *Water Gain* region

Green Dots (Water Loss):

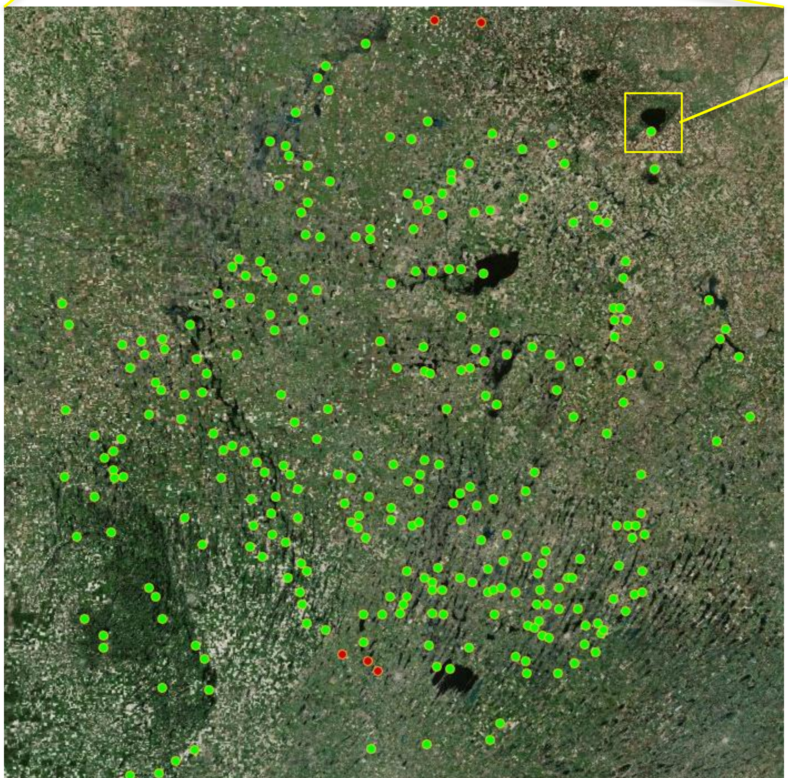
Region of size > 2.5 km^2 that have changed from water to land in the last 15 years



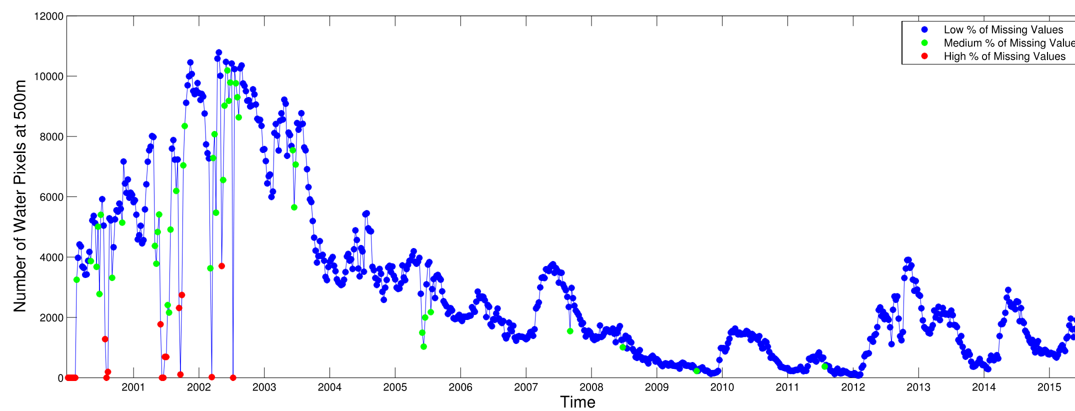
Example time series of a *Water Loss* region

Examples of Change: Shrinking Water Bodies

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

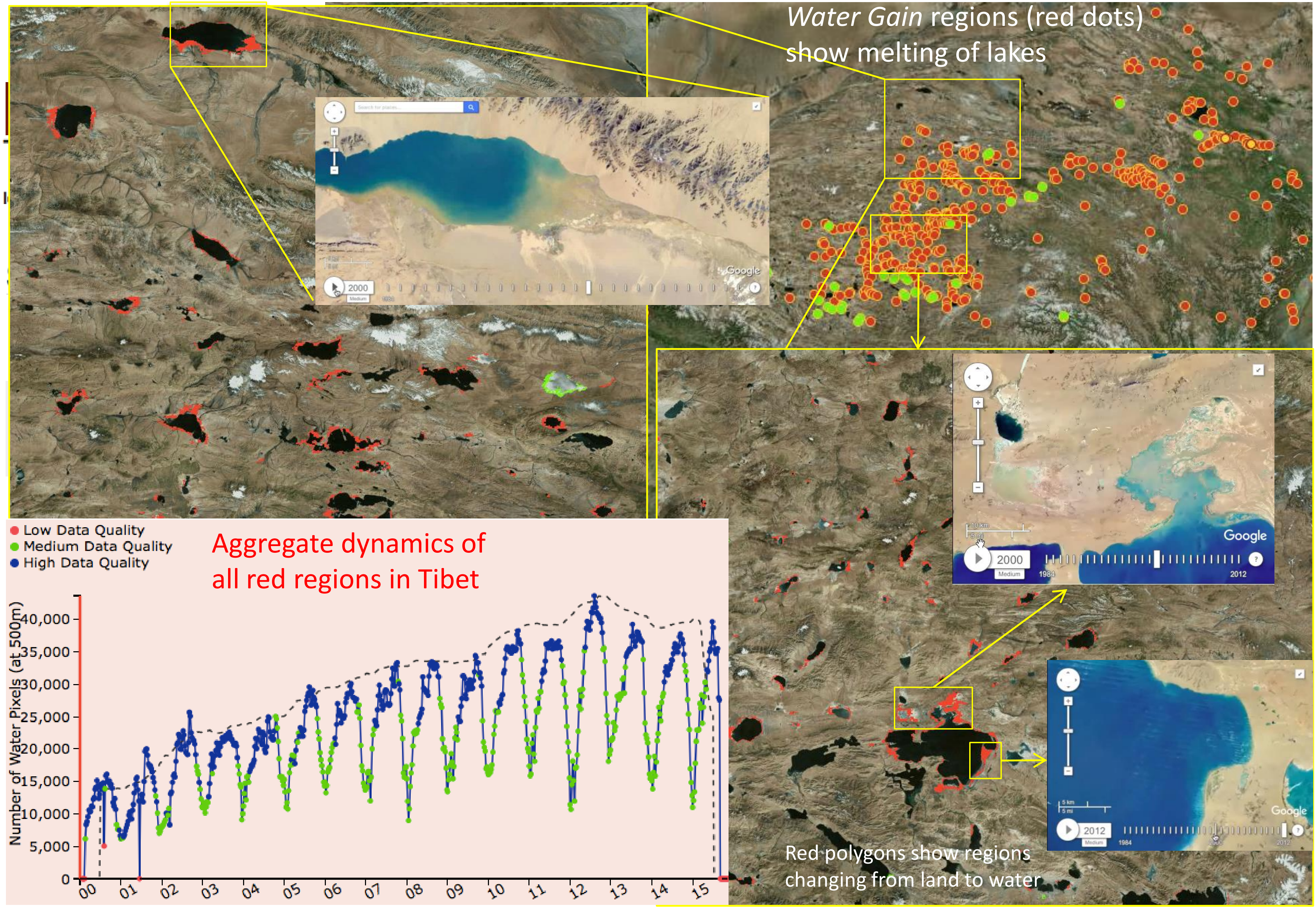


Annual Time-lapse of an example green dot



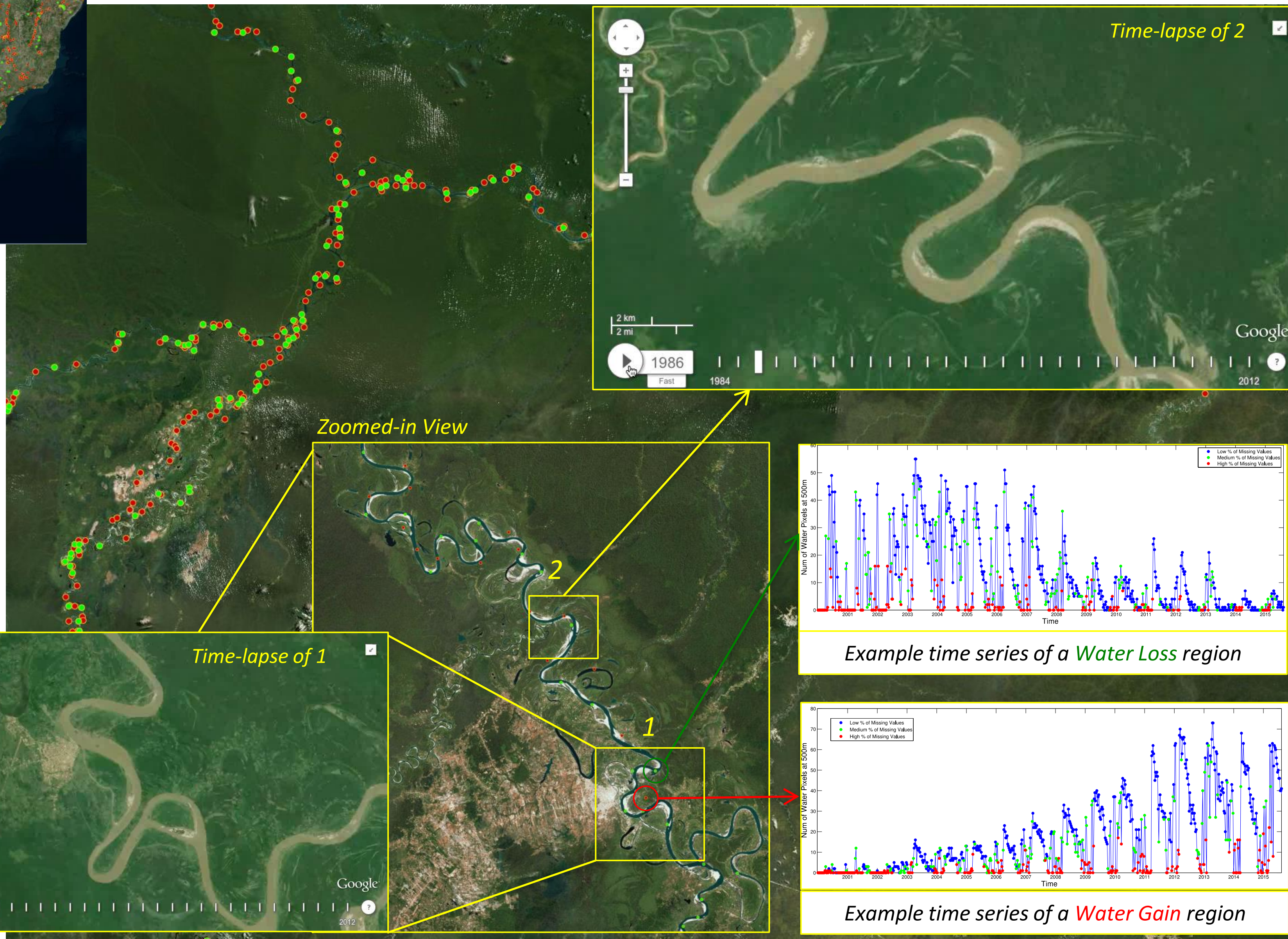
Aggregate dynamics of all green dots shown on left

Examples of Change: Melting Glacial Lakes in Tibet



Examples of Change: River Meandering

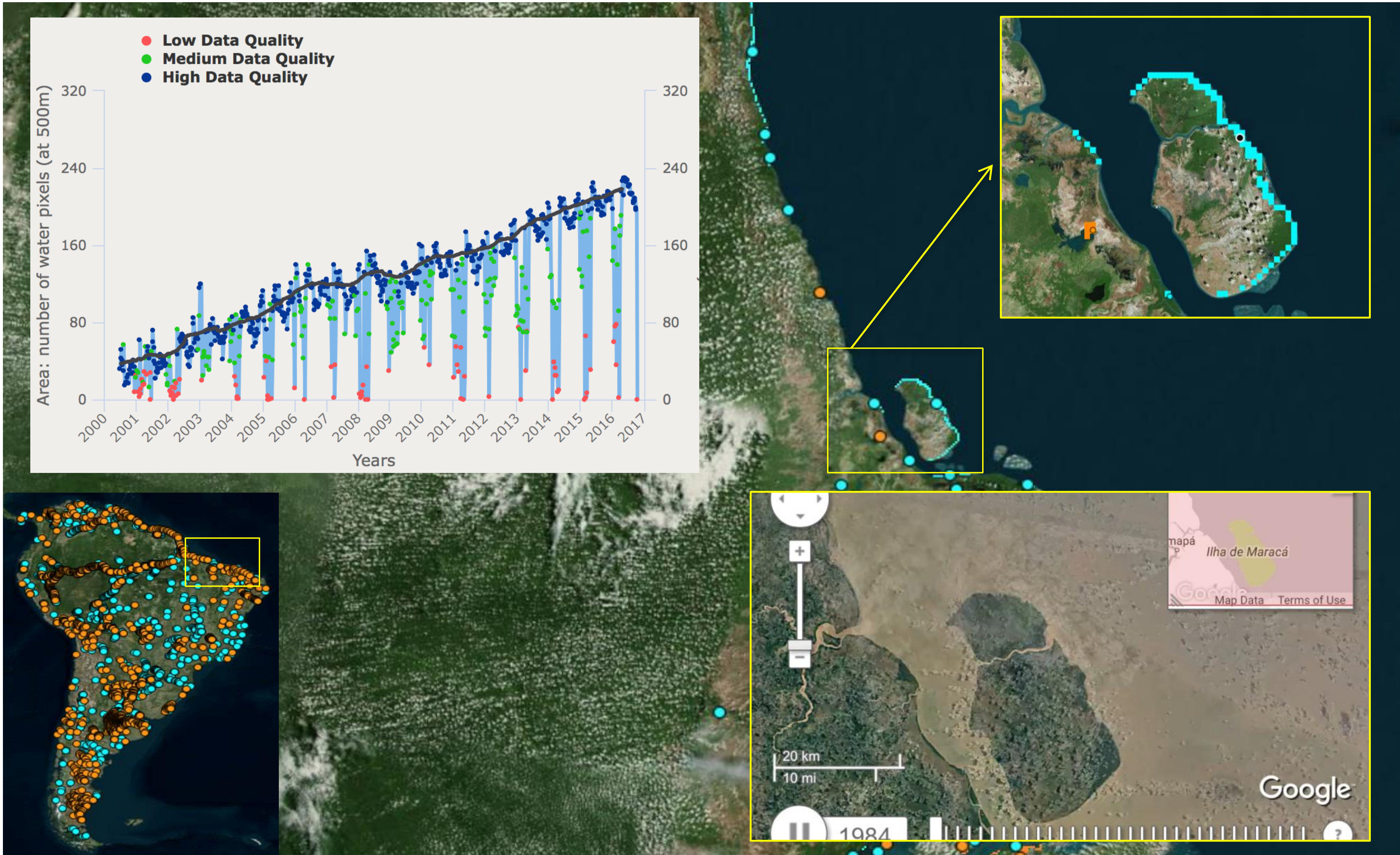
(Adjacent occurrence of *Water Gain (red)* and *Water Loss (green)* regions all along the river indicate the displacement of water from the green dots to the red dots)



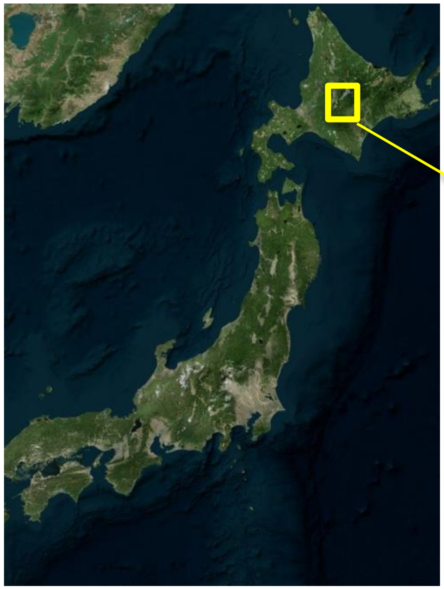
Example time series of a Water Loss region

Example time series of a Water Gain region

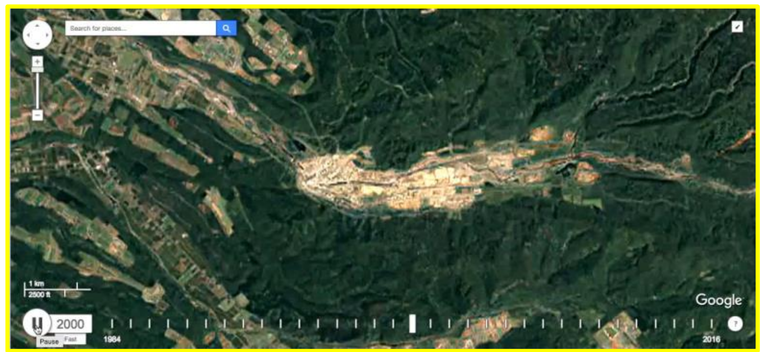
Examples of Change: Shrinking Island



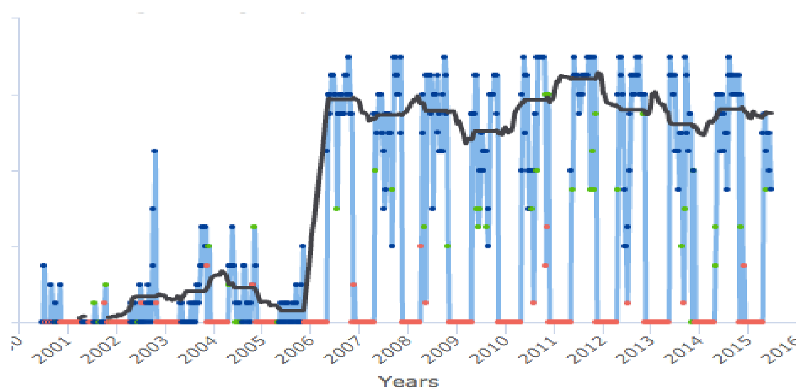
Examples of Change: Dam Construction



Construction of Chubetsu Dam, Japan



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



Global Reservoir and Dam (GRanD) Database

A data curation initiative by Global Water System Project (GWSP)

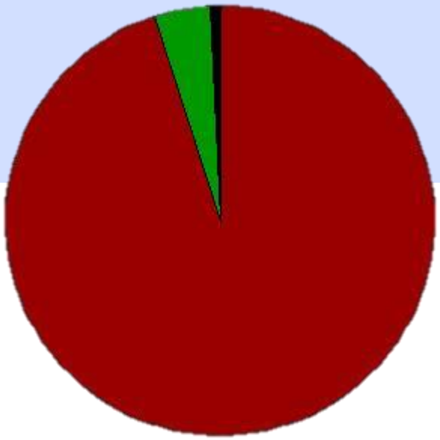
Dams reported by GRanD since 2001: 35

Comparison of Dam Detections with GRanD

Dams only reported by GRanD: 5

Dams reported by Bigdata analysis and GRanD: 30

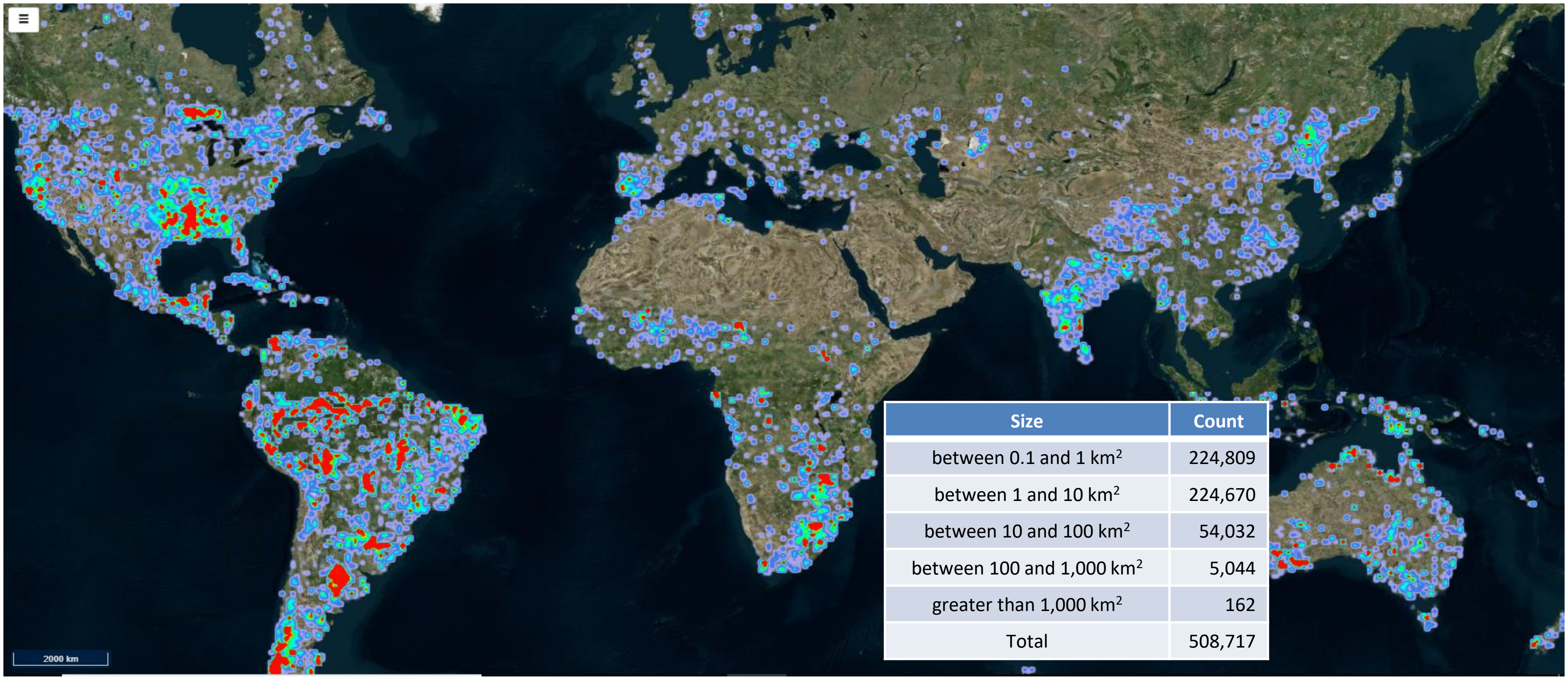
Dams only reported by Bigdata analysis: 671



Water Body Dynamics at Landsat scale using the ORBIT approach

- monthly from 1984 to 2015 (using JRC/Google land/water pixel level labels [1])
- daily from 2000 to present by downscaling daily MODIS scale land/water pixel maps

1. Pekel et. al, High-resolution mapping of global surface water and its long-term changes. Nature 540, 418-422 (2016). (doi:10.1038/nature20584)

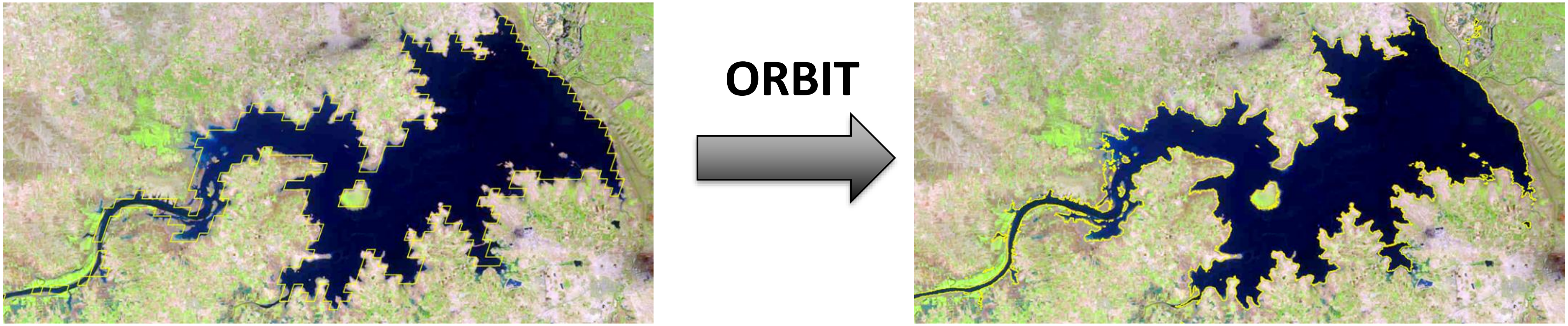


Size	Count
between 0.1 and 1 km^2	224,809
between 1 and 10 km^2	224,670
between 10 and 100 km^2	54,032
between 100 and 1,000 km^2	5,044
greater than 1,000 km^2	162
Total	508,717

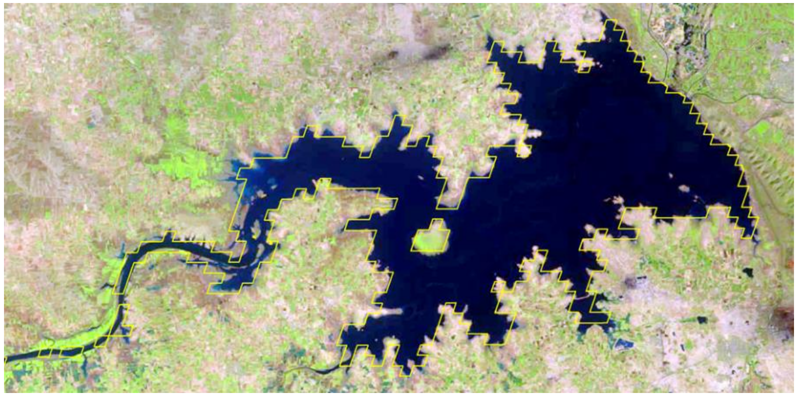
Can we produce daily surface water extents maps at high spatial resolution ?

- **Challenge:**
  - MODIS (500m resolution, **daily**)
  - LANDSAT (**30m**, every 16 days), Sentinel-2 (**10m**, every 5-10 days)

- **Solution: ORBIT - Ordering Based Information Transfer** across space and time



Can we produce daily surface water extents maps at high spatial resolution ?



ORBIT

8×8 input

32×32 samples

Super Resolution

(artificially enlarging a low resolution photograph to recover a corresponding plausible image with higher resolution)

Dahl, Ryan, Mohammad Norouzi, and Jonathon Shlens. "Pixel recursive super resolution." *arXiv preprint arXiv:1702.00783*(2017).

Concluding Remarks

- Big data in climate offers great opportunity to increase our understanding of the Earth's climate and environment as well as advance machine learning.

- Novel approaches are needed that can guide the process of knowledge discovery in scientific applications
  - "Theory-guided Data Science"

- Methods have applicability across diverse domains:
  - Ecosystem management
  - Epidemiology
  - Geospatial Intelligence
  - Neuroscience