Big Data in Climate and Earth Sciences:

Opportunities and Challenges for Machine Learning

Vipin Kumar

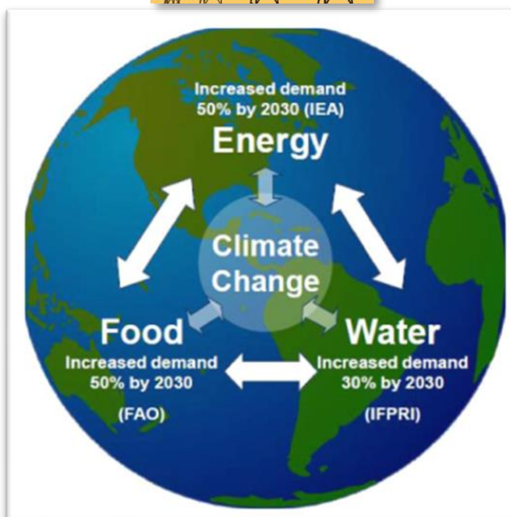
University of Minnesota

kumar001@umn.edu
www.cs.umn.edu/~kumar

Environmental Grand Challenges of the 21st Century

IPCC Report warns of ‘irreversible’ impacts of global warming *2/28/2022*

Increasing frequency of natural disasters

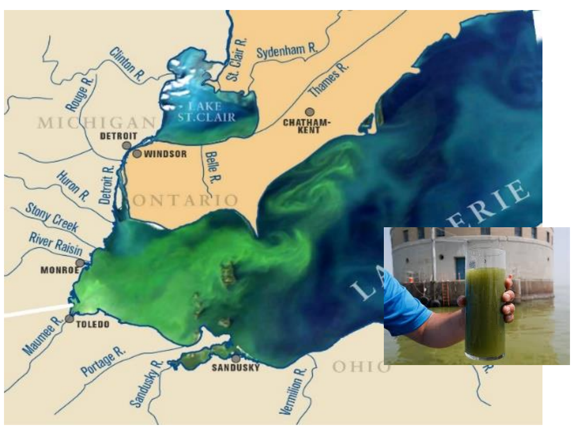


How to Feed the World Without destroying the Planet?

Cool Green Science by Nature, July 7, 2017

Oil Palm Plantations in Indonesia

Water quality impacted by Agriculture

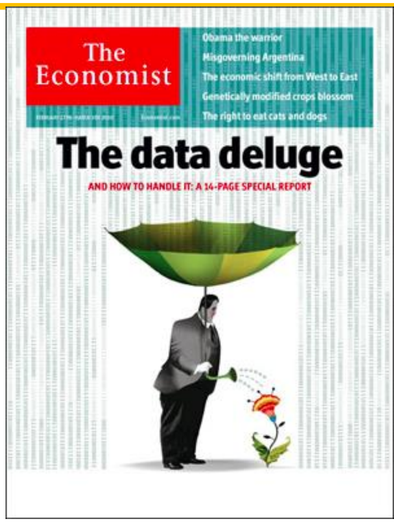


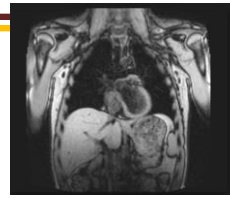
Harmful Algal Bloom in Lake Erie

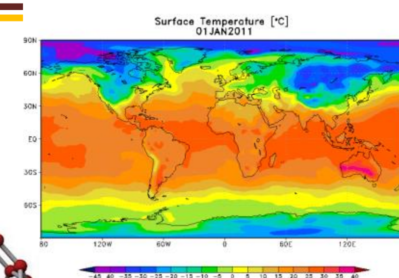
Freshwater resources under stress

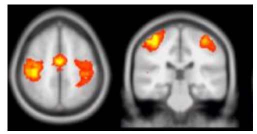
Aral Sea in 1989 Aral Sea in 2014

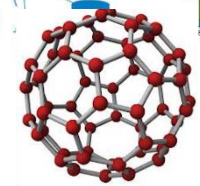
Golden Age of Data Science

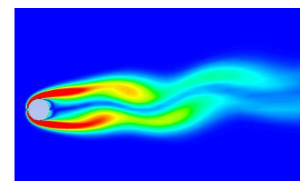












- Hugely successful in commercial applications:

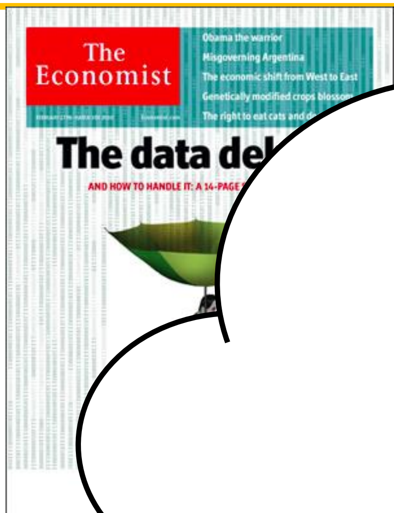
DeepMind

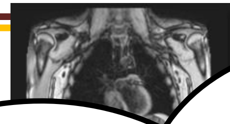
Google Ads

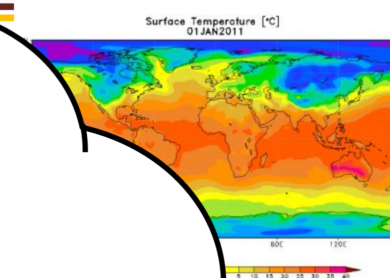
facebook

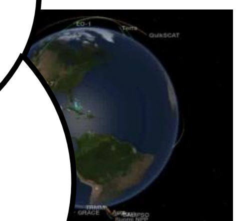
NETFLIX

Golden Age of Data Science









Challenges:

- Sparsity of labels
- Uncertain/noisy data
- Spatial and temporal heterogeneity
- Phenomena of interest can be rare
- Patterns evolvoing in space and time

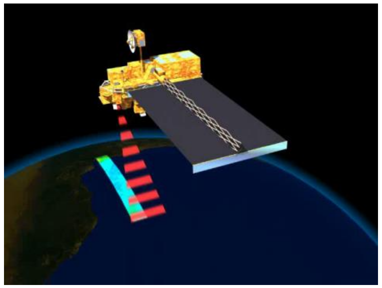
- H

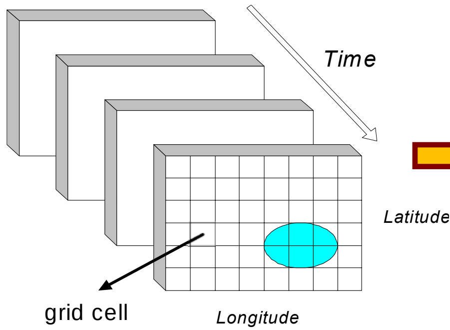
Google Ads

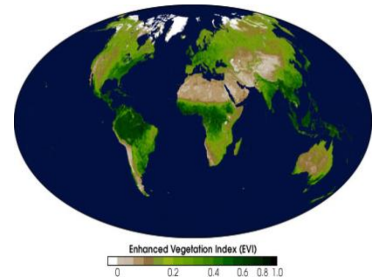
facebook

NETFLIX

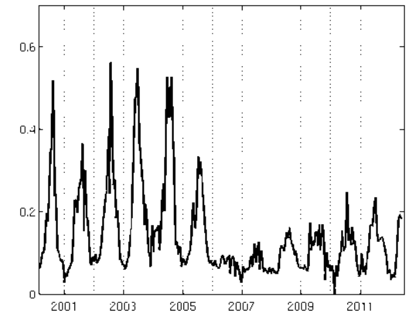
Big Data in Earth System Monitoring







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 (TKDE 2017, Remote Sensing 2018)

2. Mapping of plantation dynamics in tropical forests:

- Recurrent Neural Networks to model space and time (IEEE Big Data 2016, SDM 2016, KDD 2017, Remote Sensing 2019)

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)

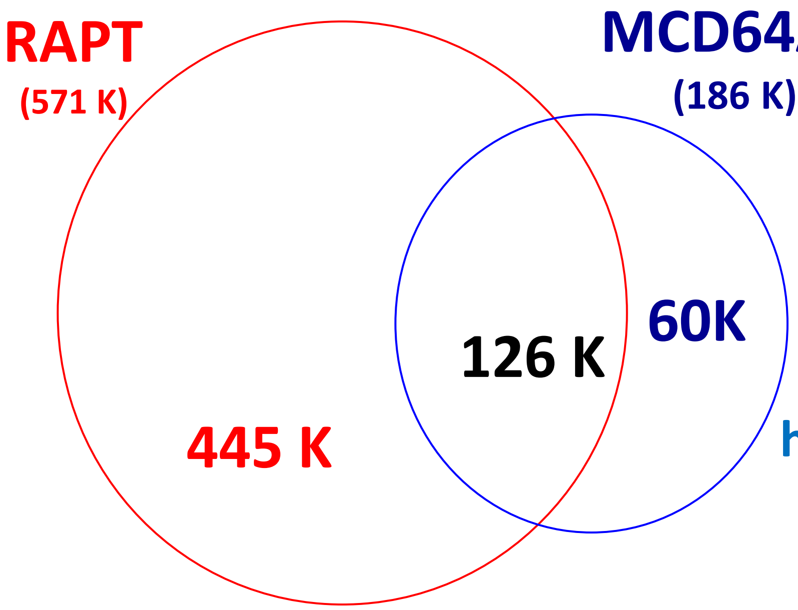
Lake Oroville in 2011 and 2014

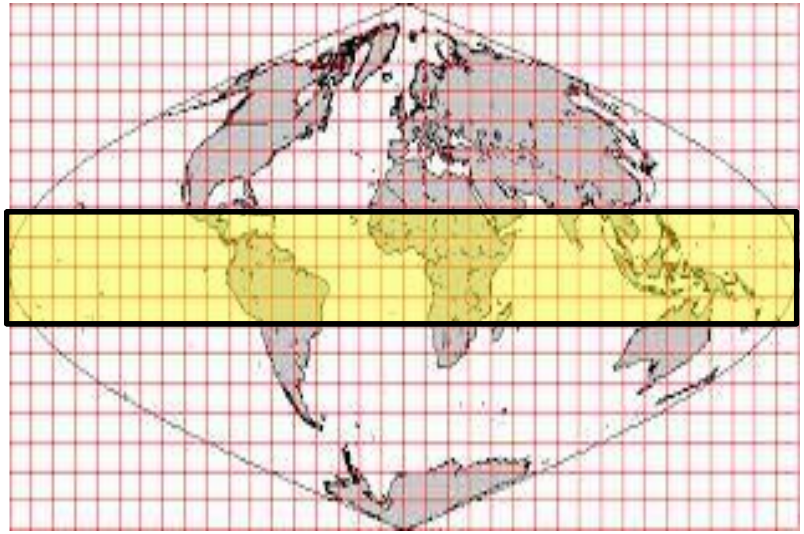
Global Monitoring of Fires in Tropical Forests

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

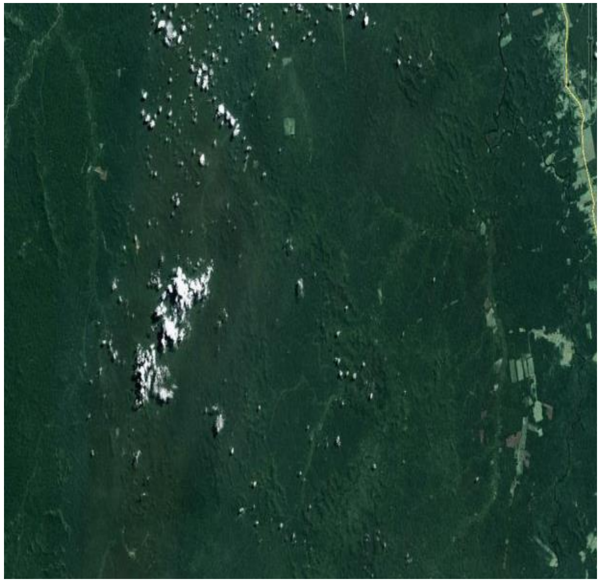




Public Viewer:
http://umnlcc.cs.umn.edu/FireMonitorRelease/

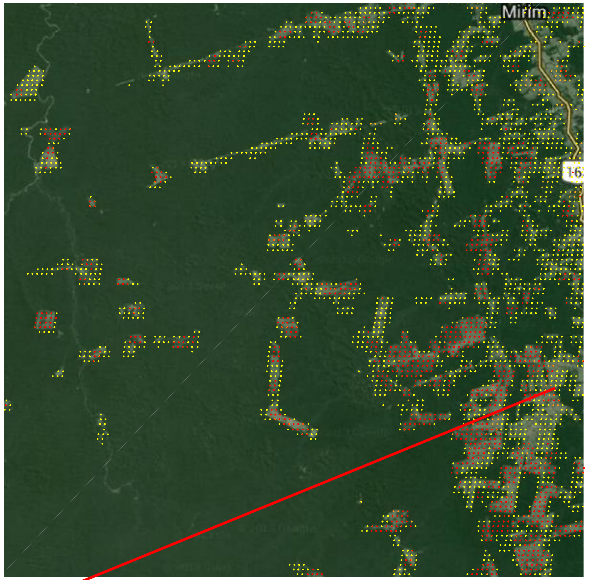
Varun Mithal et. al, "Mapping Burned Areas in Tropical Forests Using a Novel Machine Learning Framework." *Remote Sensing* 10, no. 1 (2018): 69.

Deforestation via Burning in Amazon



Google Earth Image: Year 2002

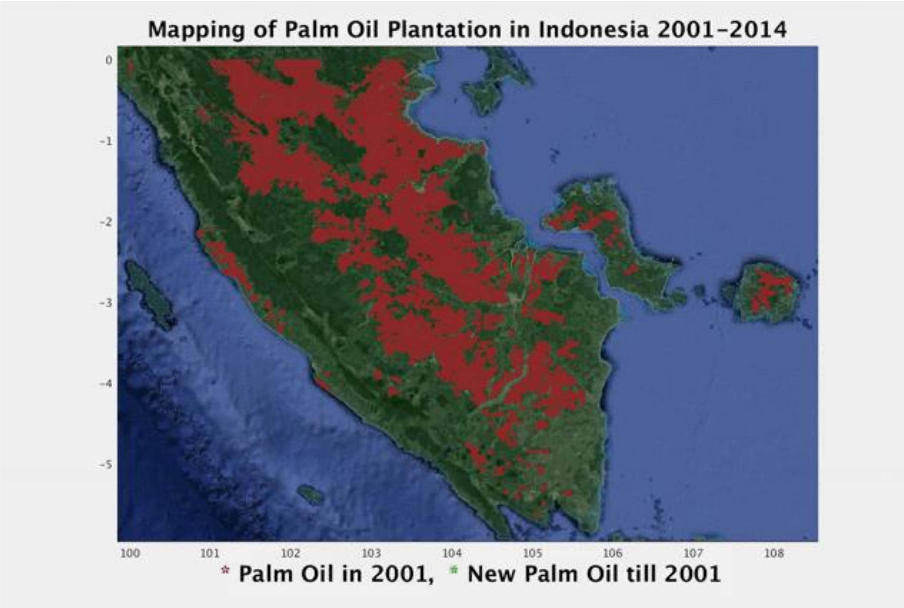
Google Earth Image: Year 2015



RAPT detection 2002-2014
(RAPT only Common)

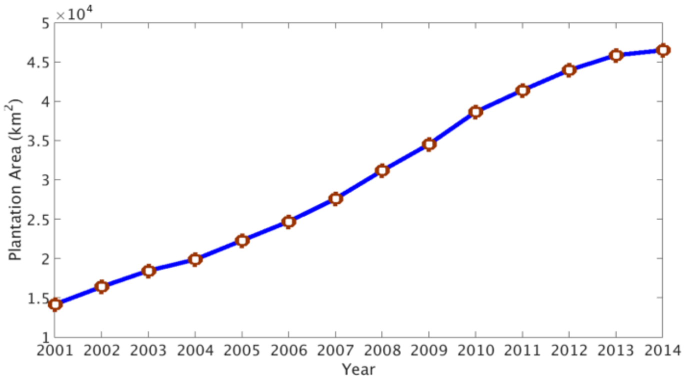
Burn Detection					**B**	**B**	**B**						
Land cover	**F**	**F**	**F**	**F**	**F**	**F**	**F**	**N**	**N**	**N**	**N**	**N**	**N**
Year	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014

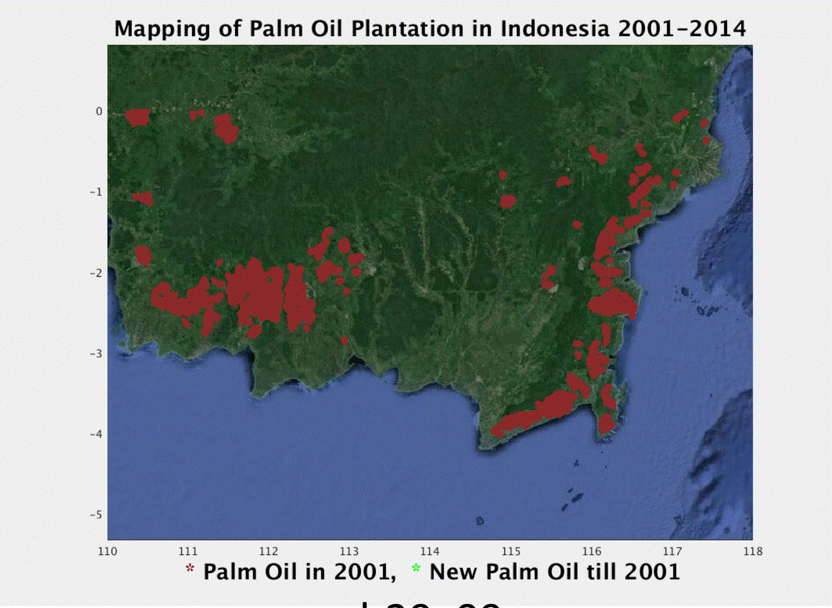
Annual Plantation Maps



h28v09

h29v08





h29v09

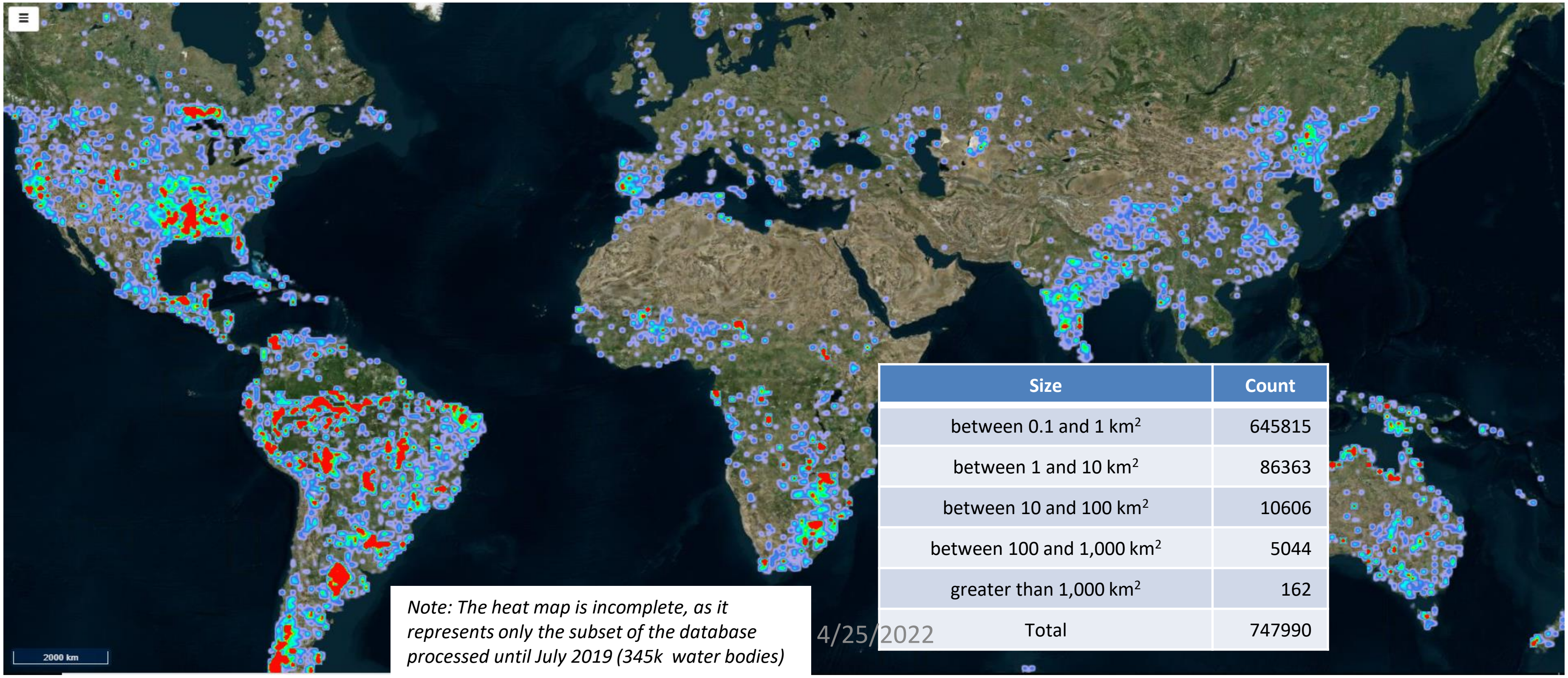
- ❑ Annual growth rate ≈ 9.57%

RealSAT: Reservoir and Lake Surface Area Time-series Database

http://umnlcc.cs.umn.edu/realsat/

- Monthly scale surface area dynamics from 1984 to 2015 at 30m resolution
  - using JRC-Google product as input label source [1]
- Over 700k water bodies of size greater than 0.1 sq. kms below 50 degrees North

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)



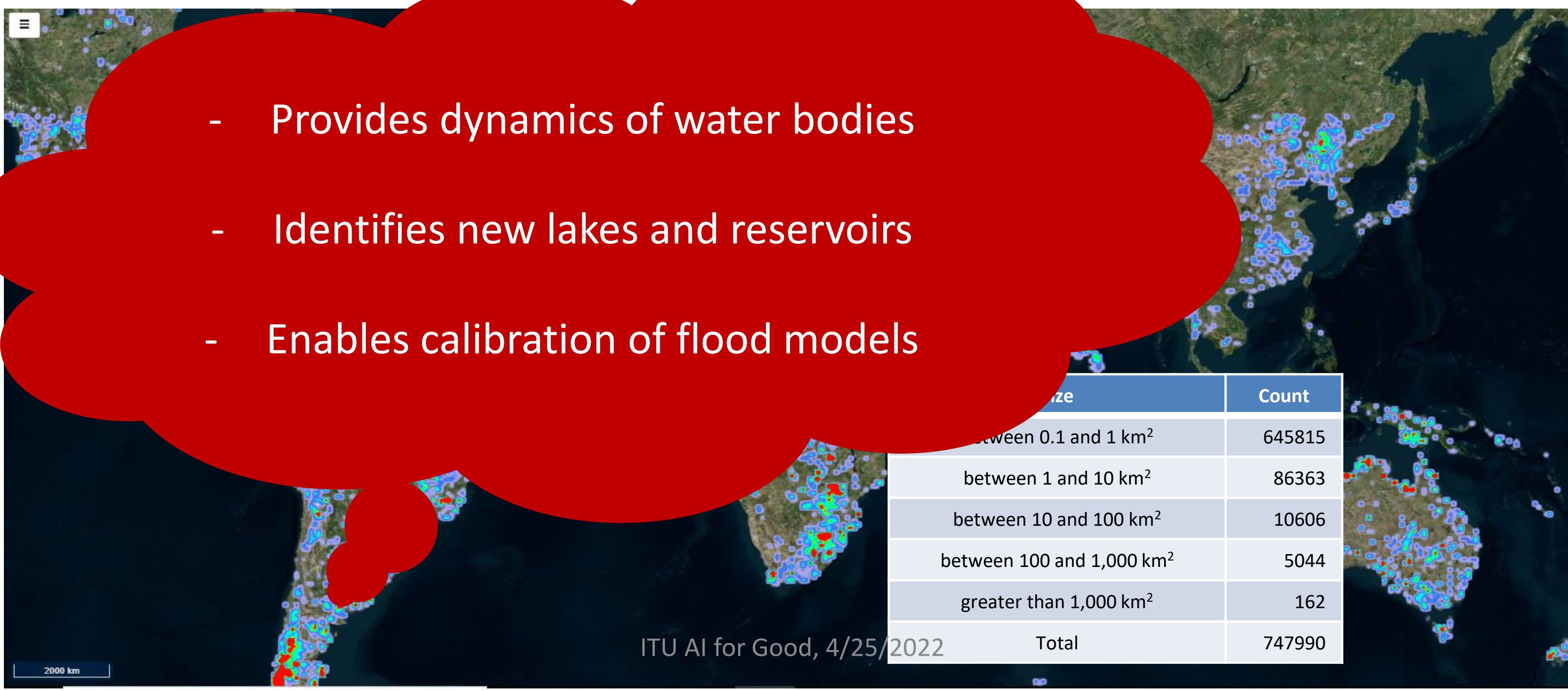
Note: The heat map is incomplete, as it represents only the subset of the database processed until July 2019 (345k water bodies)

Size	Count
between 0.1 and 1 km^2	645815
between 1 and 10 km^2	86363
between 10 and 100 km^2	10606
between 100 and 1,000 km^2	5044
greater than 1,000 km^2	162
Total	747990

ReaLSAT: Resrvoir and Lake Surface Area Timeseries Database
http://umnlcc.cs.umn.edu/realsat/

- Monthly scale dynamics from 1984 to 2015 at 30m resolution
  - using JRC-Google product as input label source [1]
- Over 700k water bodies of size greater than 0.1 sq. kms.

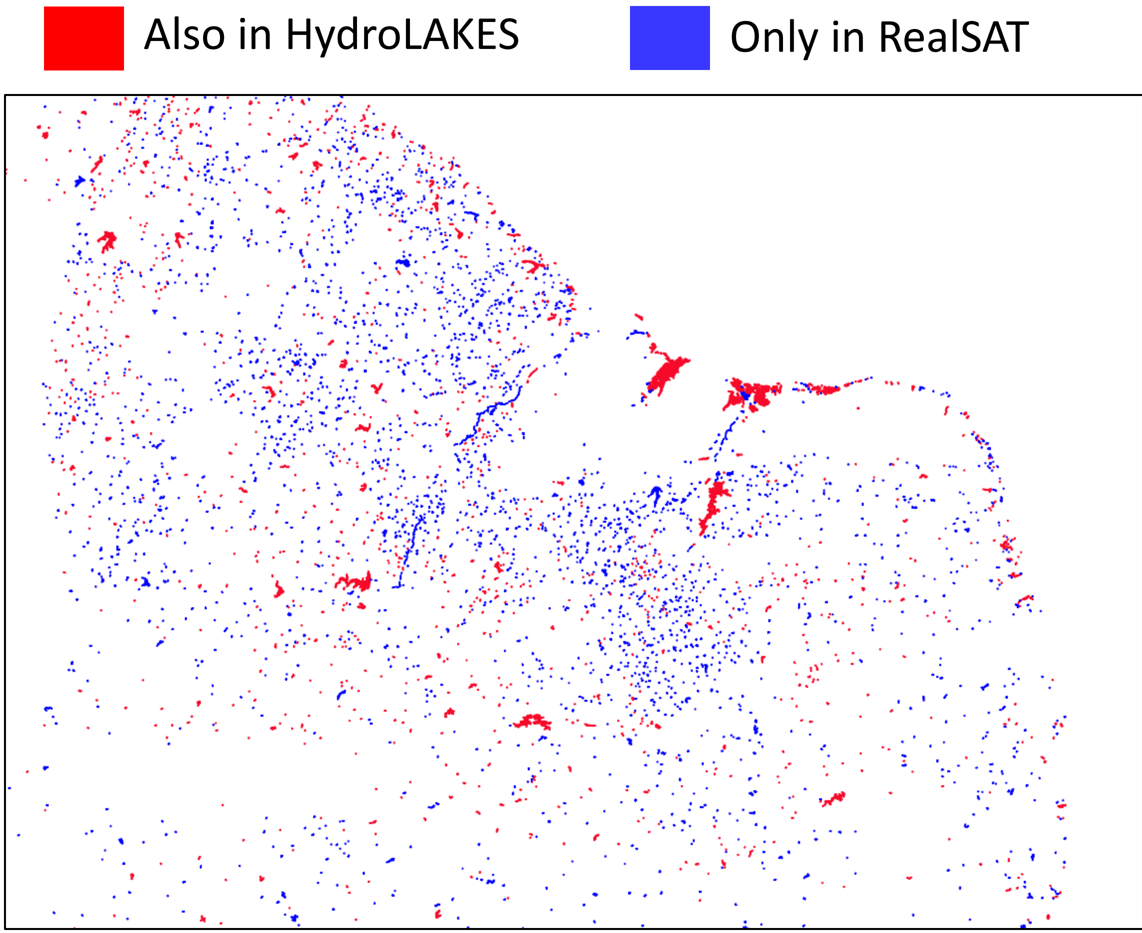
1. Pekel et. al, High-resolution mapping of glob... water an... ange... -422 (2016). (doi:10.1038/nature20584)



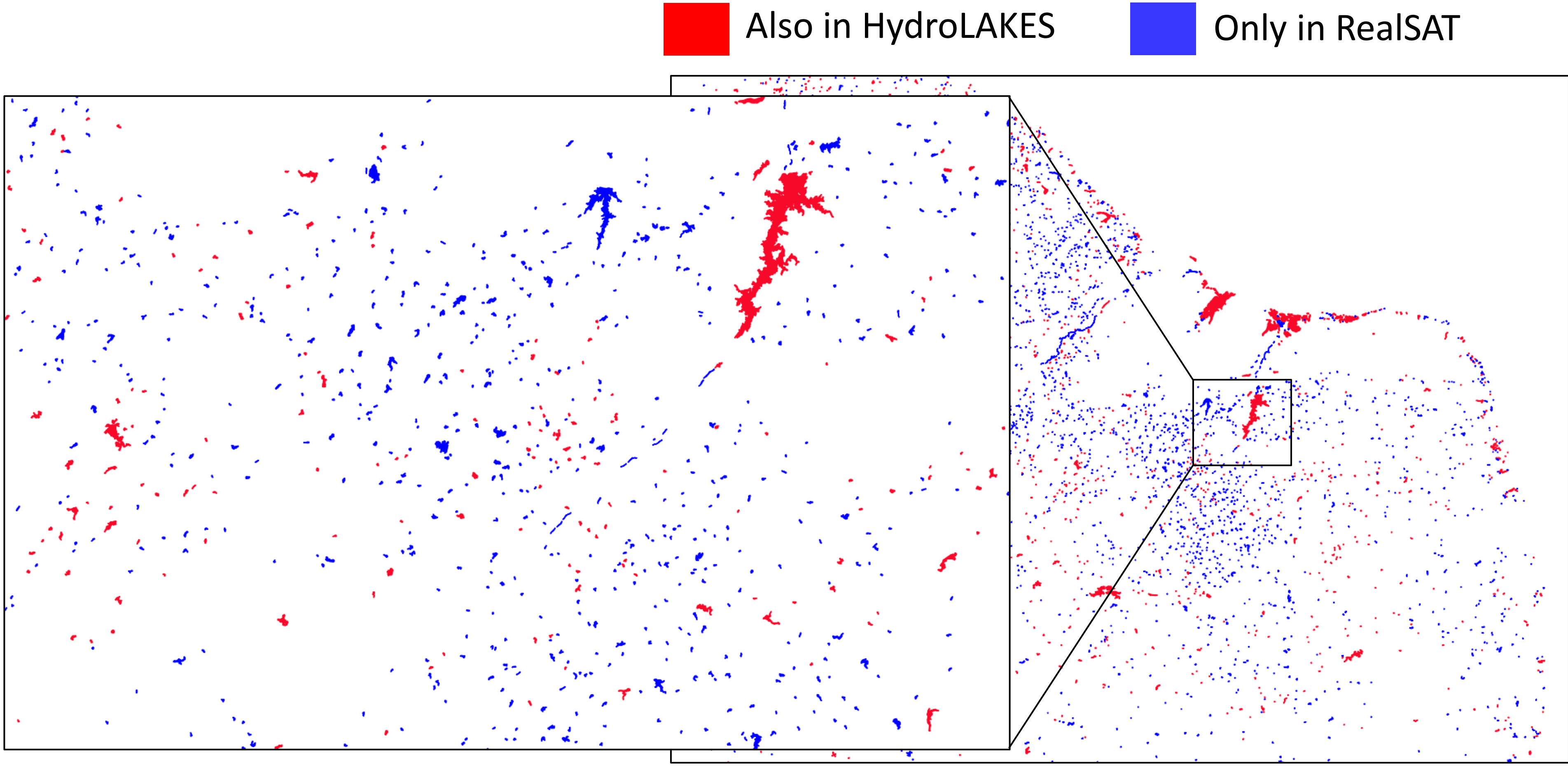
- Provides dynamics of water bodies
- Identifies new lakes and reservoirs
- Enables calibration of flood models

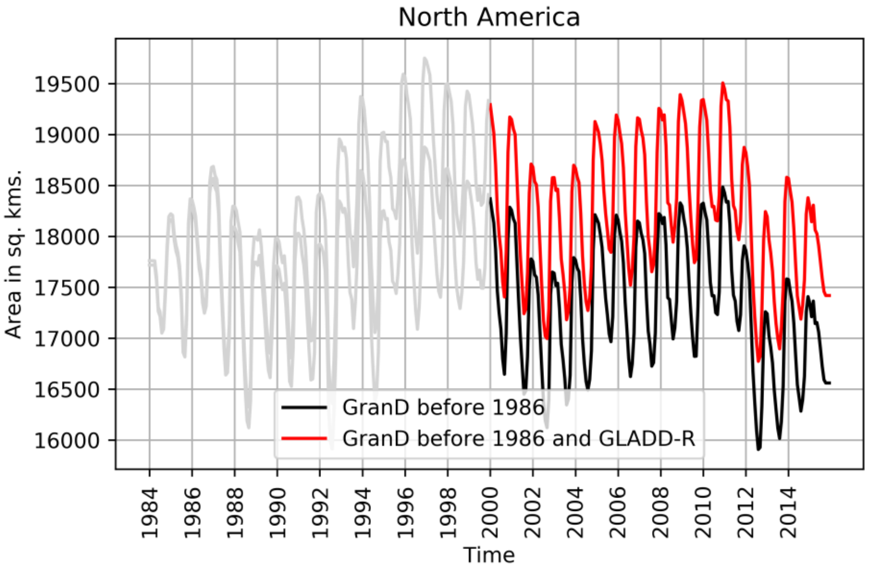
...ze	Count
...tween 0.1 and 1 km^2	645815
between 1 and 10 km^2	86363
between 10 and 100 km^2	10606
between 100 and 1,000 km^2	5044
greater than 1,000 km^2	162
Total	747990

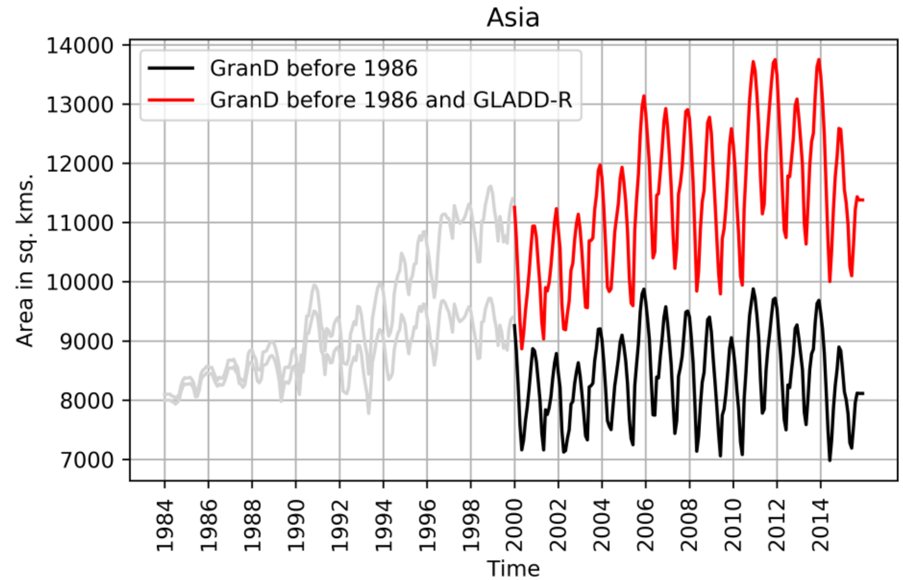
Water bodies in South America

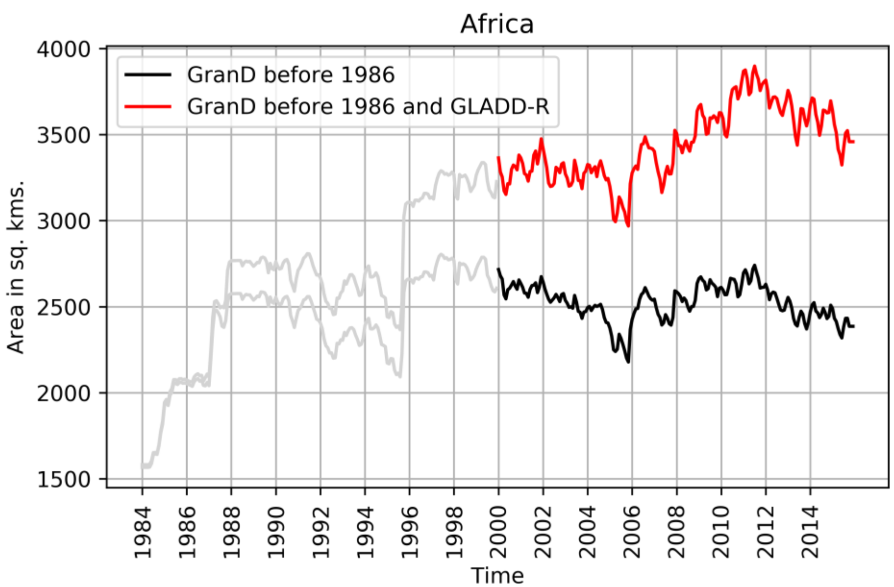


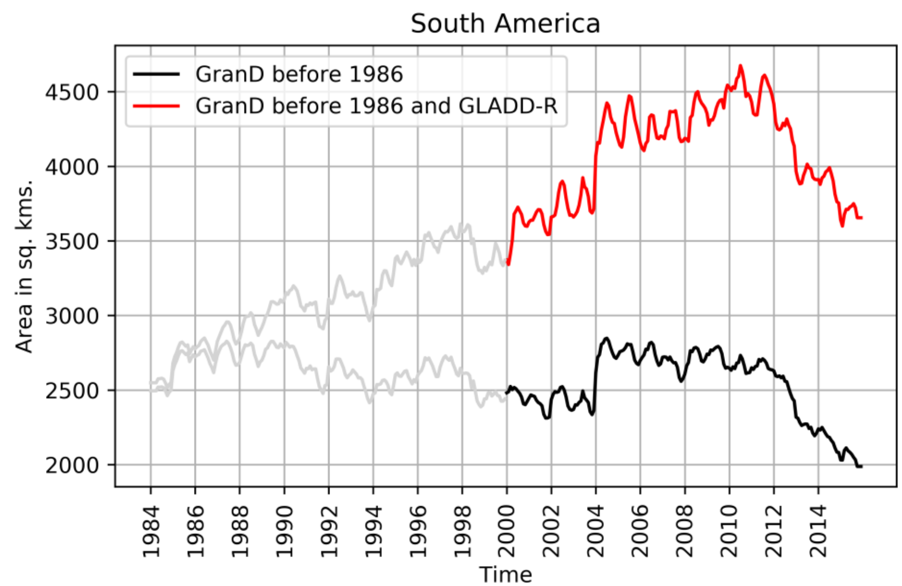
Water bodies in South America



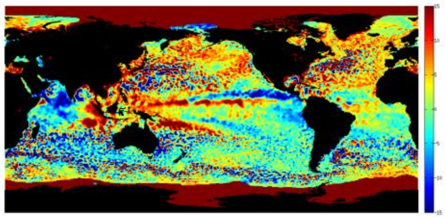

North America
GranD before 1986
GranD before 1986 and GLADD-R
Area in sq. kms.
Time


Asia
GranD before 1986
GranD before 1986 and GLADD-R
Area in sq. kms.
Time


Africa
GranD before 1986
GranD before 1986 and GLADD-R
Area in sq. kms.
Time


South America
GranD before 1986
GranD before 1986 and GLADD-R
Area in sq. kms.
Time

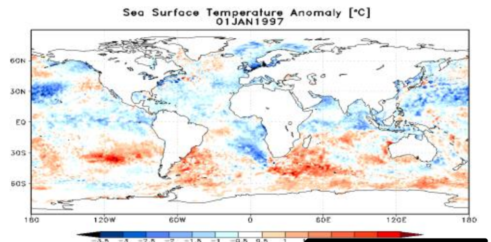
ML for Environmental Sciences: Additional Research Highlights



Pattern Mining

Monitoring Ocean Eddies

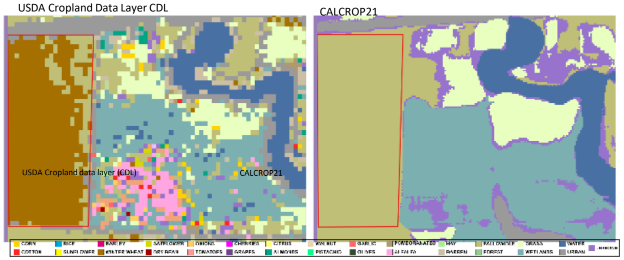
- Spatio-temporal pattern mining using novel multiple object tracking algorithms
- Created an open source data base of 20+ years of eddies and eddy tracks



Network Analysis

Climate Teleconnections

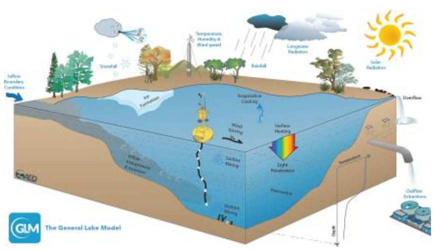
- Scalable method for discovering related graph regions
- Discovery of novel climate teleconnections
- Also applicable in analyzing brain fMRI data



Spatio-Temporal Semantic Segmentation

Mapping Crops at Scale

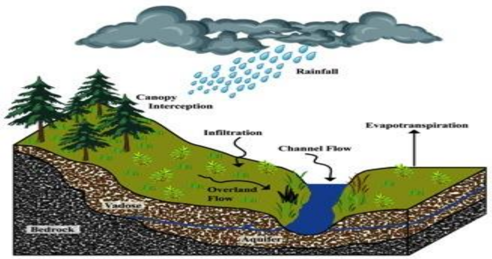
- Attention augmented deep learning algorithm jointly exploits the spatial and temporal nature of satellite data.
- Created a 10m resolution crop map for the entire California Crop Belt that is more accurate than the 30m resolution CDL used as labels for training



Knowledge Guided Machine Learning (KGML)

Modeling Lake Water Quality

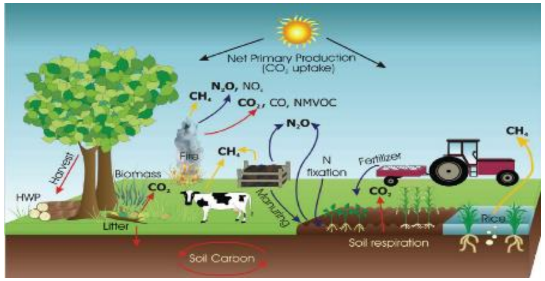
- Combining Physics and Data Science models to overcome complementary weaknesses
- Hybrid models are significantly more robust and are of higher quality than either pure physics or pure data models



Self-supervised KGML

Inverse Modelling in Hydrology

- Deep learning based inverse framework for estimating features given driver-response data
- Proposed framework reduces uncertainty in the catchment characteristics of hydrological basins.



KGML for Agro-Ecosystem Sustainability

Modelling GHG emission in Agriculture

- Scientific Knowledge guided process based models
- KGML-ag provides much more accurate modeling of GHG emissions relative to state of the art process based models

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

Understanding Climate Change: A Data-driven Approach

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

IPCC INTERGOVERNMENTAL PANEL ON CLIMATE CHANGE

WCRP World Climate Research Programme

WMO UNEP

WORLD ECONOMIC FORUM

Climate Science Special Report

Fourth National Climate Assessment (NCA4), Volume I

computing in SCIENCE & ENGINEERING — Computing & Climate — IEEE

WORLD ECONOMIC FORUM — Agenda | Initiatives | Reports | Events | About

How can we better manage our planet's water supply? 25 Sep 2015

Nonlinear Processes in Geophysics

National Science Foundation WHERE DISCOVERIES BEGIN

Home › News | Discovery

Using data to better understand climate change

NSF-supported research team develops data-driven methods to refine climate predictions, analyze climatic changes August 23, 2016

National Science Foundation WHERE DISCOVERIES BEGIN

Home › News | Discovery July 30, 2014

Climate change research goes to the extremes

science360 NEWS — BREAKING SCIENCE NEWS THAT SHAPES YOUR WORLD

TOP STORY

Researchers Devise More Accurate Method For Predicting Hurricane Activity

NASA Landsat Science

The Secret Lives of Migrating Rivers — News — Dec 16, 2016

Highlights:

- **150+ publications** in high-profile computer science and geoscience venues
  - Papers in Nature and Nature Climate Change
  - Best paper awards, e.g., SDM, KDD, …
- **Education and Outreach Activities:**
  - Nurturing Climate Informatics community
  - Special Issues: CiSE, Nonlinear Geophysics
  - Sessions at SDM, ICDM, KDD, AGU, AMS
- **Involvement in National/International Assessments:**
  - US 4th National Climate Assessment
  - 5th Assessment Report of the IPCC
- **Collaboration with Federal/International Agencies:**
  - NASA, NCAR, USGS, DOE, DHS/FEMA
  - WCRP, UNEP, WMO, IPCC
- **Industry Collaborations and Spinoffs:**
  - Startups: risQ, Planetary Skin, …
  - Insurance: AIR, AIG, Swiss Re, Tokio Marine,…

risQ

Pioneering Innovative Climate Data Research

nature climate change

Lack of uniform trends but increasing spatial variability in observed Indian rainfall extremes

A daily global mesoscale ocean eddy dataset from satellite altimetry

www.nature.com/ SCIENTIFIC DATA

nature International weekly journal of science

NATURE | RESEARCH HIGHLIGHTS

Climate change: Cold spells in a warm world

nature International weekly journal of science

NATURE | NEWS

How machine learning could help to improve climate forecasts Nicola Jones 23 August 2017

nature International weekly journal of science

NATURE | LETTER

Intensification and spatial homogenization of coastal upwelling under climate change

Concluding Remarks

- Big data and machine learnng offers great opportunity to increase our understanding of the Earth’s climate and environment

- Addressing challenges unique to environmental problems require methodological advances in 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

Team Members

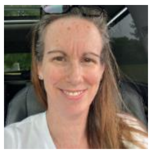

Saurabh Agrawal


Gowtham Atluri


Shyam Boriah


Ivan Brugere

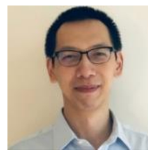

Xi Chen


Kelly Cutler


James Faghmous


Ashish Garg


Rahul Ghosh


Xiaowei Jia


Anuj Karpatne


Jaya Kawale


Ankush Khandelwal


Arjun Kumar


Lydia Manikonda


Varun Mithal


Guruprasad Nayak


Praveen Ravirathinam

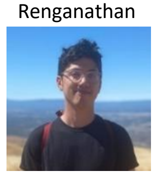

Arvind Renganathan


Somya Sharma


Michael Steinbach


Kshitij Tayal


Karsten Steinhaueser


Jared Willard


Shaoming Xu

Acknowledgements

NSF: 1029711, Expeditions in Computing: Collaborative Research: Understanding Climate Change: A Data Driven Approach; 1739191, INFEWS/T3: Innovations for Sustainable Food, Energy, And Water Supplies In Intensively Cultivated Regions: Integrating Technologies, Data, And Human Behavior; 1838159, BIGDATA:F: Advancing Deep Learning to Monitor Global Change; 1943721, HDR Collaborative Research: Knowledge Guided Machine Learning: A Framework for Accelerating Scientific Discovery;201962, STC: Learning the Earth with AI and Physics
NASA: NNX16AB21G, Scalable Analysis of Earth System Data Using Parallelized Graph-Based Approaches; NNX12AP37G, Integrating Parallel and Distributed Data Mining Algorithms into the NASA Earth Exchange (NEX); **Planetary Skin Institute**: Global Land Use Change; **USGS:** Process guided machine learning for water temperature prediction**; ARPA-E**: SMARTFARM Phase 2

Collaborators

UMN: Arindam Banerjee, Snigdhansu Chatterjee, Zhenong Jin, Stefan Liess, John Nieber, Phil Pardey, Shashi Shekhar

NCSU: Nagiza Samatova, Fredrick Semazzi
Northeastern University: Auroop Ganguly
North Carolina A&T: Abdollah Homaifar
NASA Ames: Ramakrishna Nemani, Nikunj Oza
Penn State: Christopher Duffy
UCLA: Dennis Lettenmaier, Miriam Marlier
U of Wisconsin: Paul Hanson, Hilary Dugan
USGS: Alison Appling, Jordan Read, Jeff Stadler, Jacob Zwartz