Geodata Education at Purdue

Wen-wen Tung
Earth, Atmospheric, and Planetary Sciences
Earth System Science Data Lab

wwtung@purdue.edu

education

- Earth, Atmospheric, and Planetary Science Geodata Professional MS Concentration (est. Spring 2018)

- Data Science and Geodata Science Graduate-level MOOCs based on EAPS509/515 and STAT695 (est. Fall 2018)

- Ongoing:

    - 3+2 BS+MS Geodata Professional MS

    - Geodata Science Undergraduate MOOCs

    - Interdisciplinary Certificate at PhD level

computing

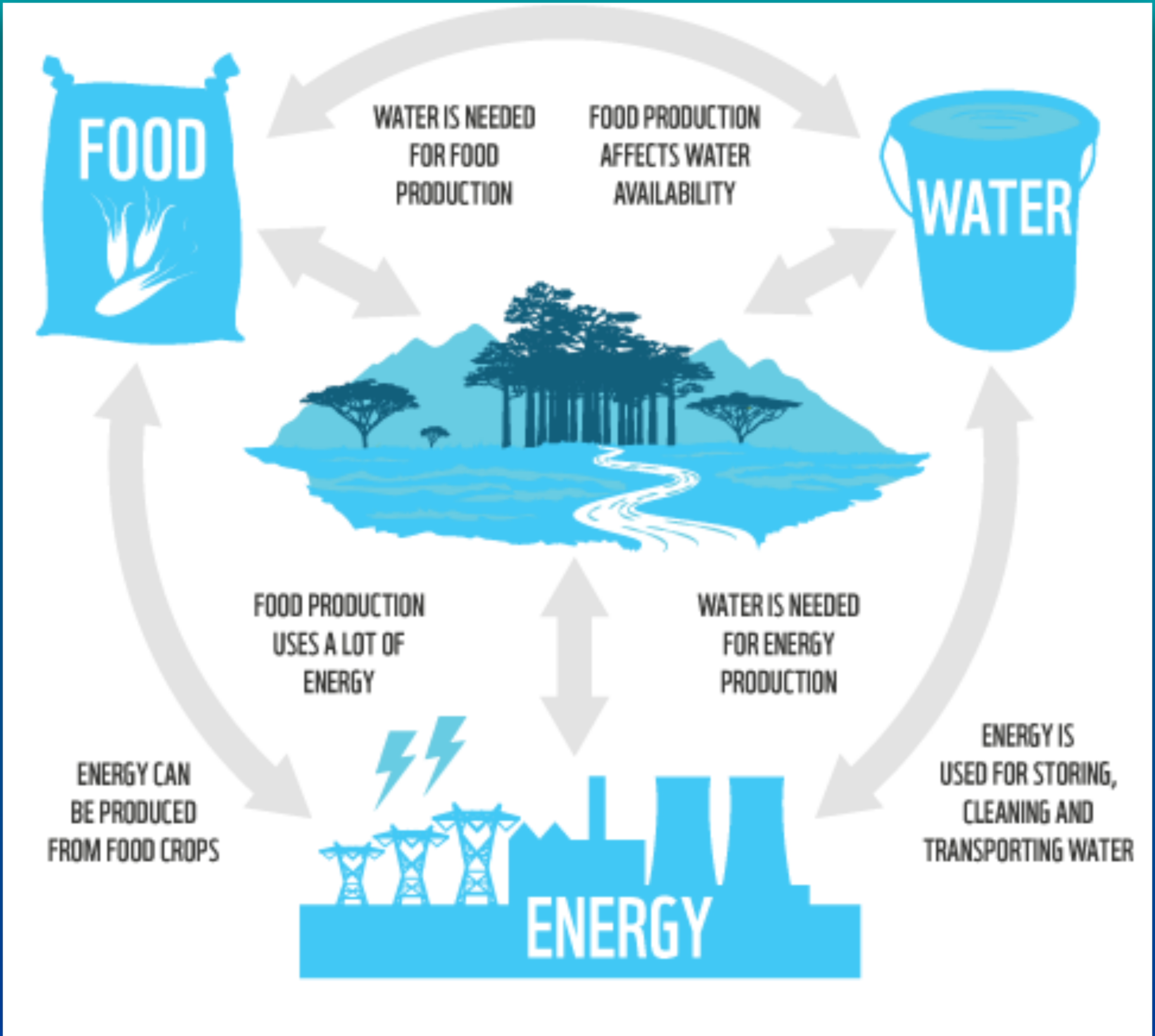
- Existing Hathi and WSC Hadoop Cluster
- New Provost-funded WCERES (Weather, Climate, Environmental, Resource, Energy, and Societal) Nexus Hadoop Cluster
  - 10 nodes + frontend + name node, 24 cores each (240 cores in total), 128 GB RAM each node, 8x4 TB each node (320 TB in total)

research support group

- WCERES Consortium (Mainly Purdue, a proposal has been submitted to NSF EarthCube for nation-wide participation)
- International Data Science Consortium

what motivated WCERES

A cluster of faculty and their students in EAPS, STAT/CS, CE, ABE, AGRY, IE, & Krannert



WATER IS NEEDED FOR FOOD PRODUCTION

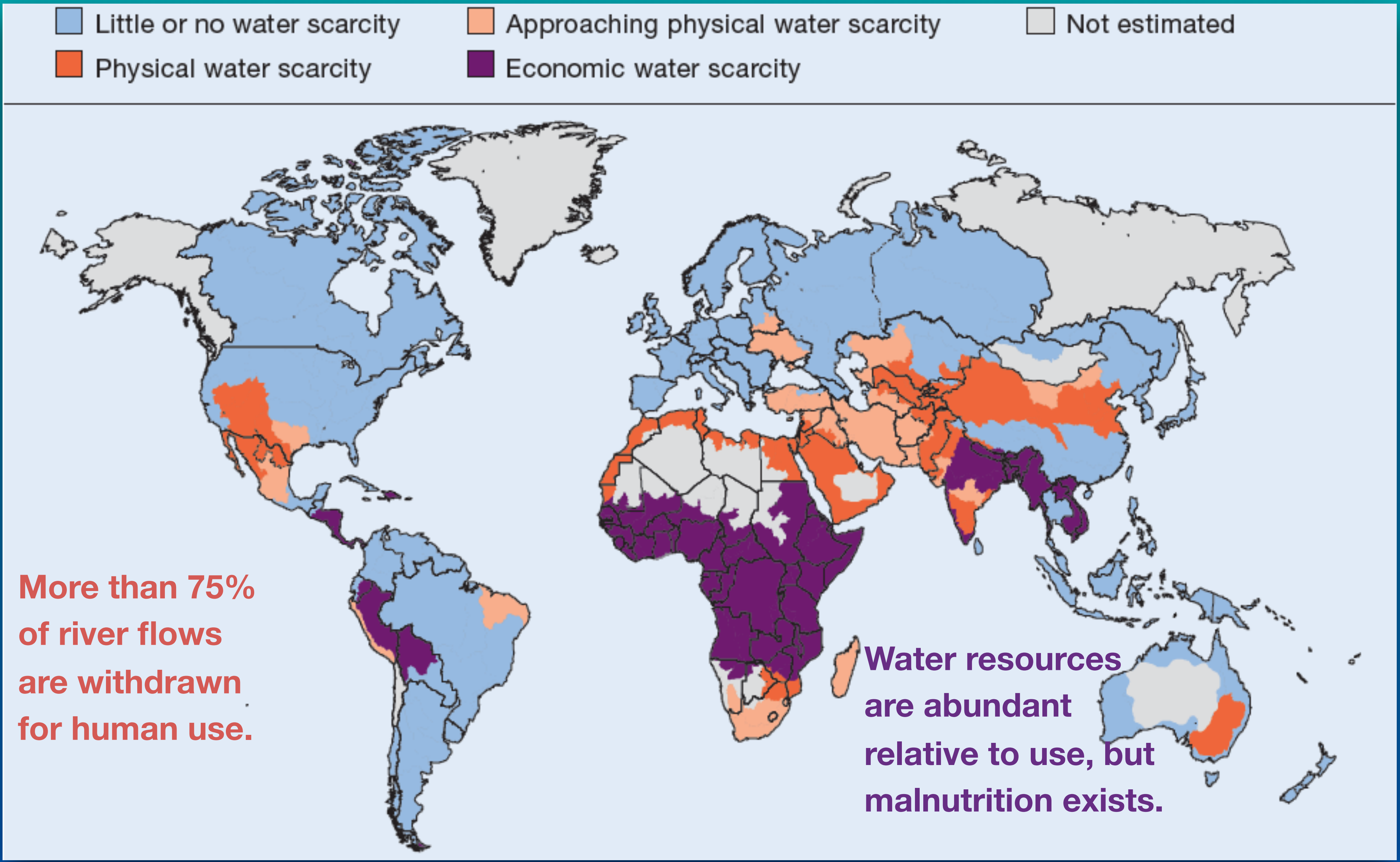
FOOD PRODUCTION AFFECTS WATER AVAILABILITY

FOOD PRODUCTION USES A LOT OF ENERGY

WATER IS NEEDED FOR ENERGY PRODUCTION

ENERGY CAN BE PRODUCED FROM FOOD CROPS

ENERGY IS USED FOR STORING, CLEANING AND TRANSPORTING WATER

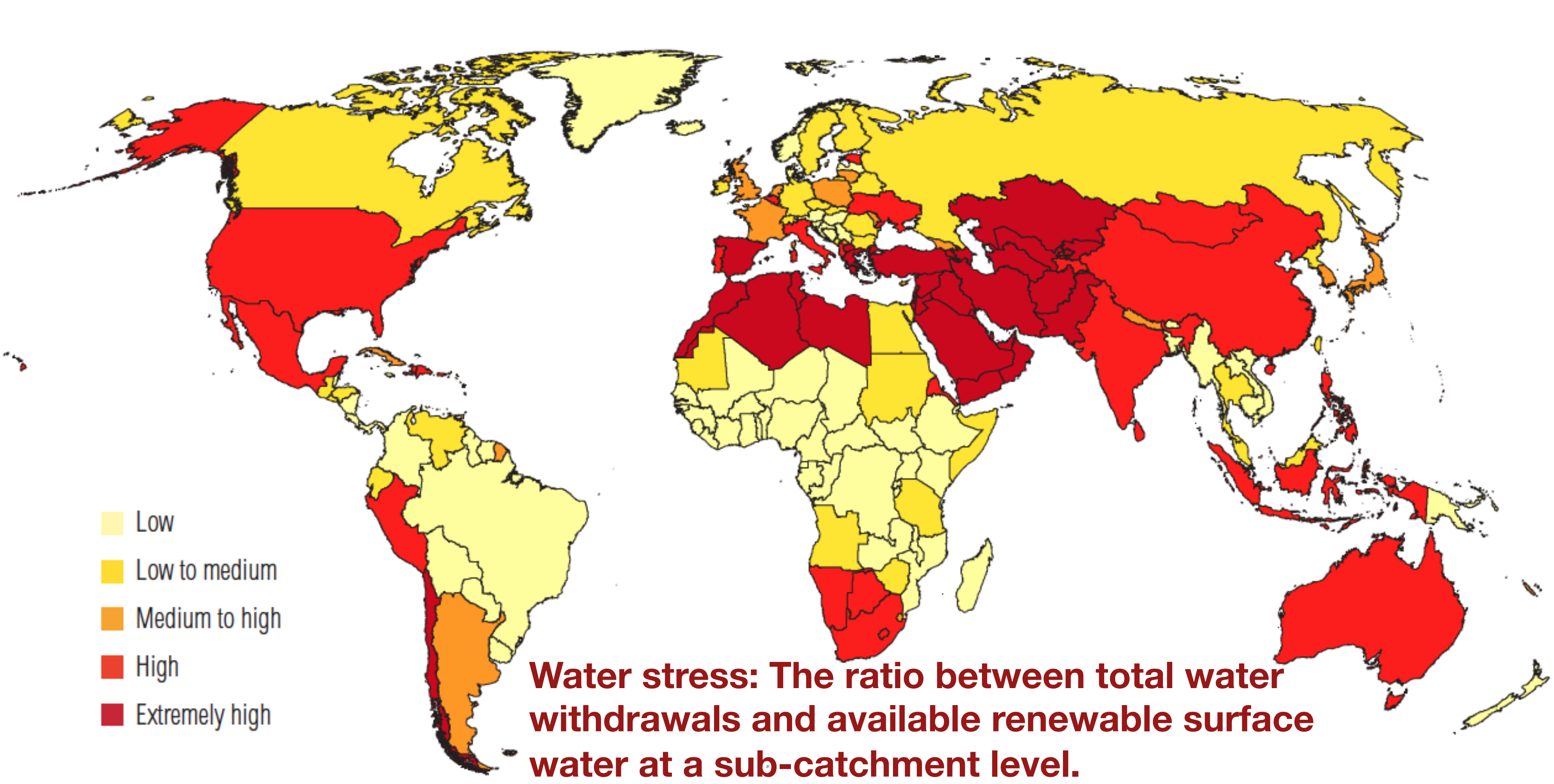

🟦 Little or no water scarcity
🟧 Approaching physical water scarcity
⬜ Not estimated
🟥 Physical water scarcity
🟪 Economic water scarcity

More than 75% of river flows are withdrawn for human use.

Water resources are abundant relative to use, but malnutrition exists.

Areas of Physical and Economic Water Scarcity

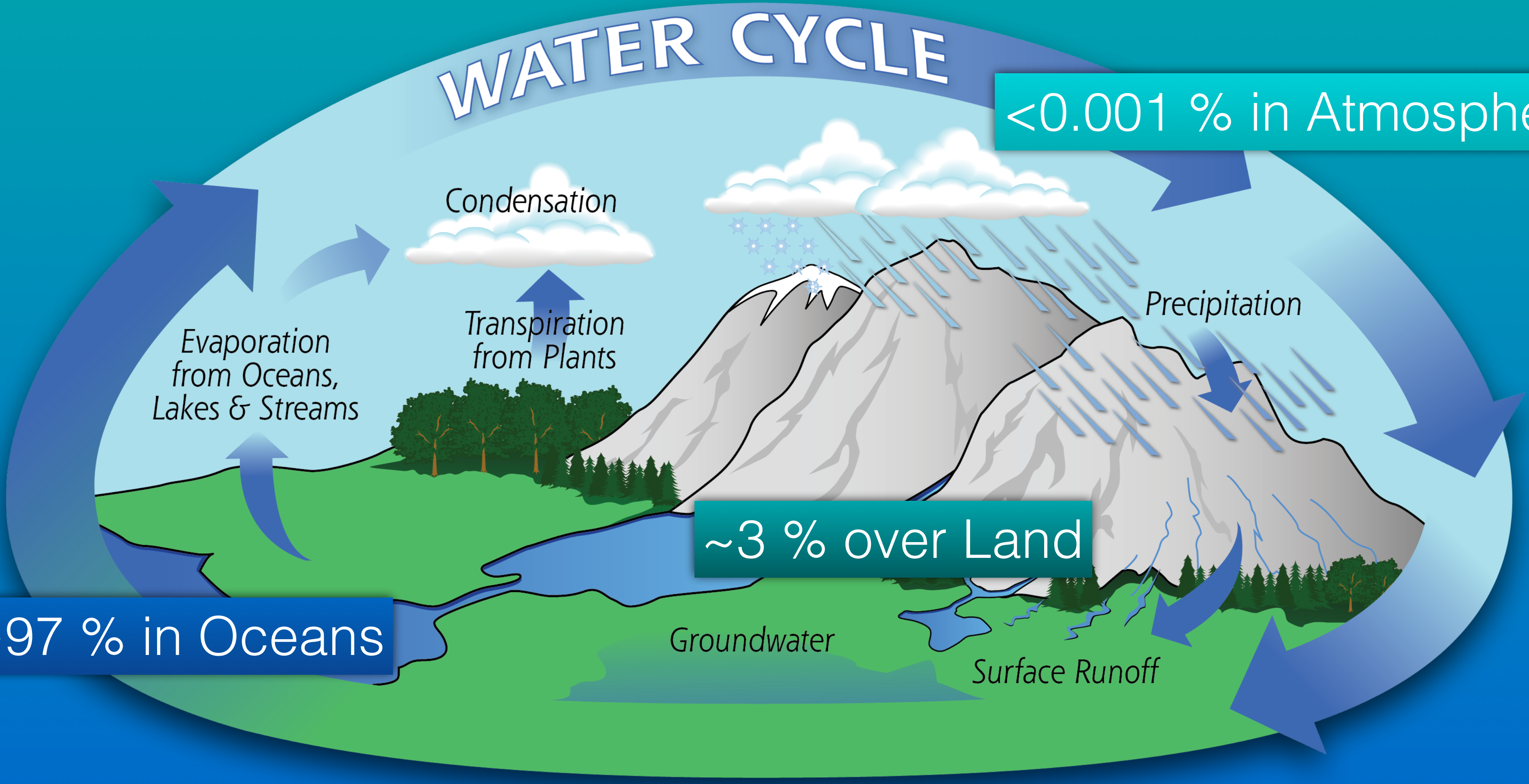
“Comprehensive Assessment of Water Management in Agriculture” (2007, International Water Management Institute)



Water stress: The ratio between total water withdrawals and available renewable surface water at a sub-catchment level.

Country-level **Water Stress** in 2040 under the Business-As-Usual Scenario

Luo et al. (2015) “Aqueduct Projected Water Stress Country Rankings”, World Resources Institute


WATER CYCLE
Condensation
Evaporation from Oceans, Lakes & Streams
Transpiration from Plants
Precipitation
Groundwater
Surface Runoff

<0.001 % in Atmosphere

~3 % over Land

~97 % in Oceans

Water is the primary medium by which matter and energy are circulated in the Earth systems;
it is central to the regional and global Security of Food, Energy, and other Resources

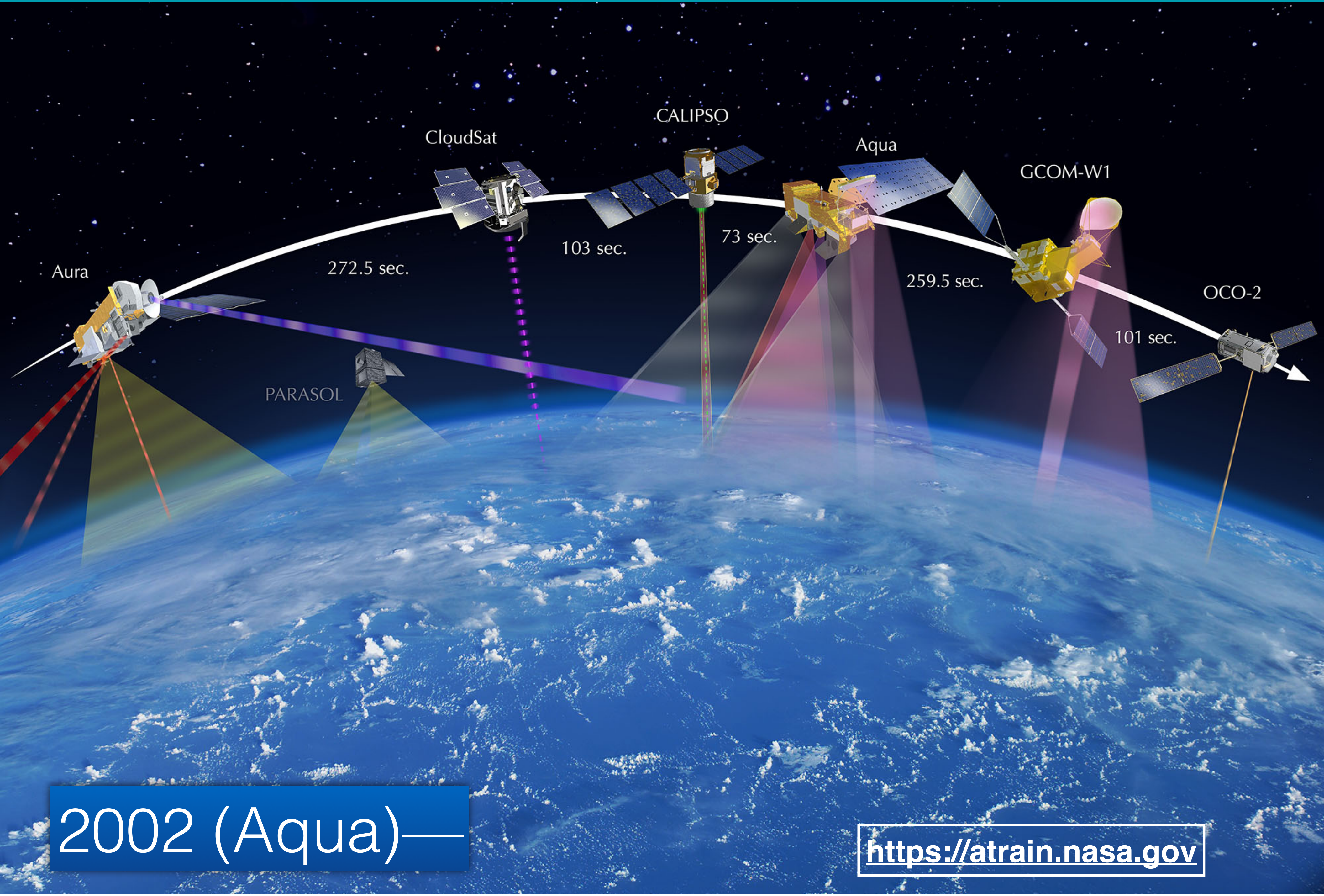
http://pmm.nasa.gov

Deep Analysis of Large Complex Satellite-Based Cloud and Precipitation Data

Wen-wen Tung[1], William S. Cleveland[2,3],
Matthew C. Bowers[1], and Wanchen Wu[1,4]

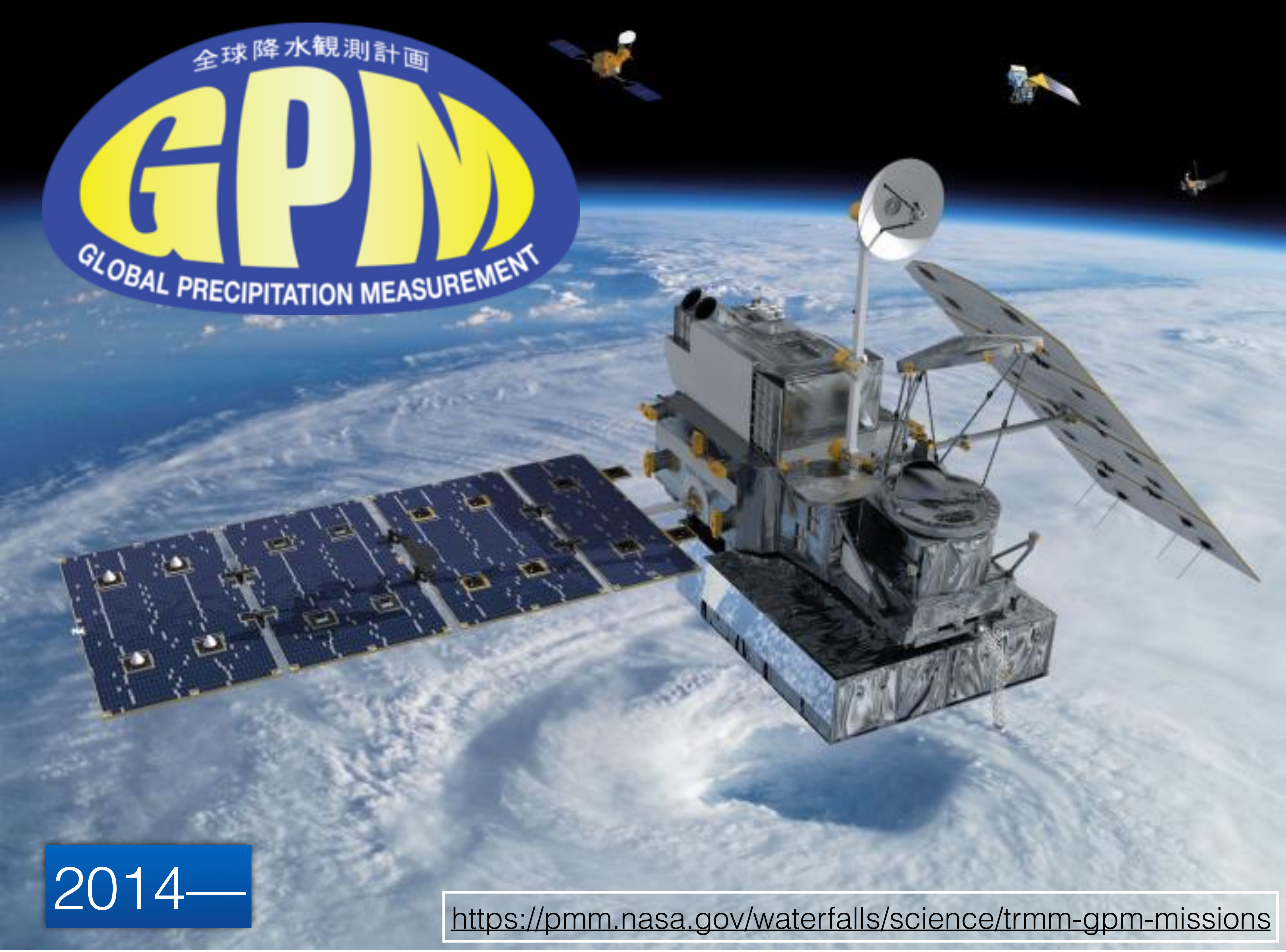
[1]Department of Earth, Atmospheric, & Planetary Sciences,
[2]Department of Statistics,
[3]Department of Computer Science, Purdue,
[4]Academia Sinica, Taipei, Taiwan

The Afternoon Constellation — A-Train



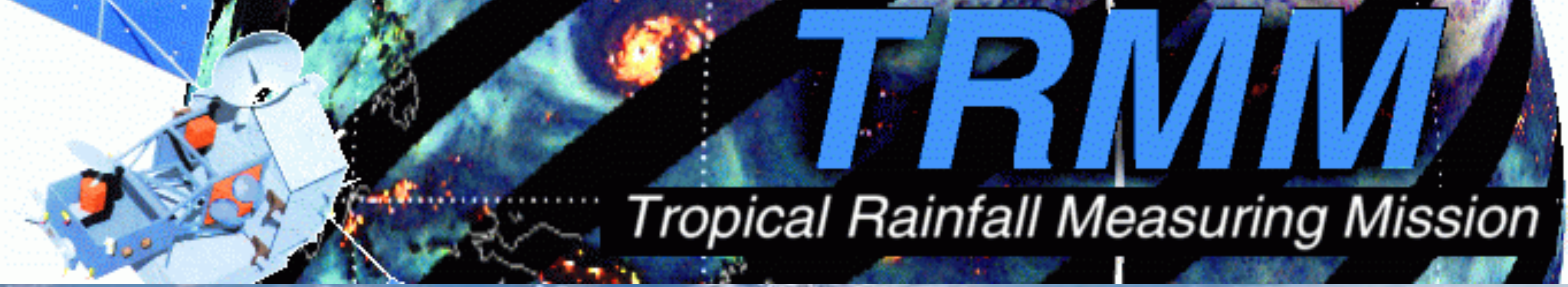
2002 (Aqua)—

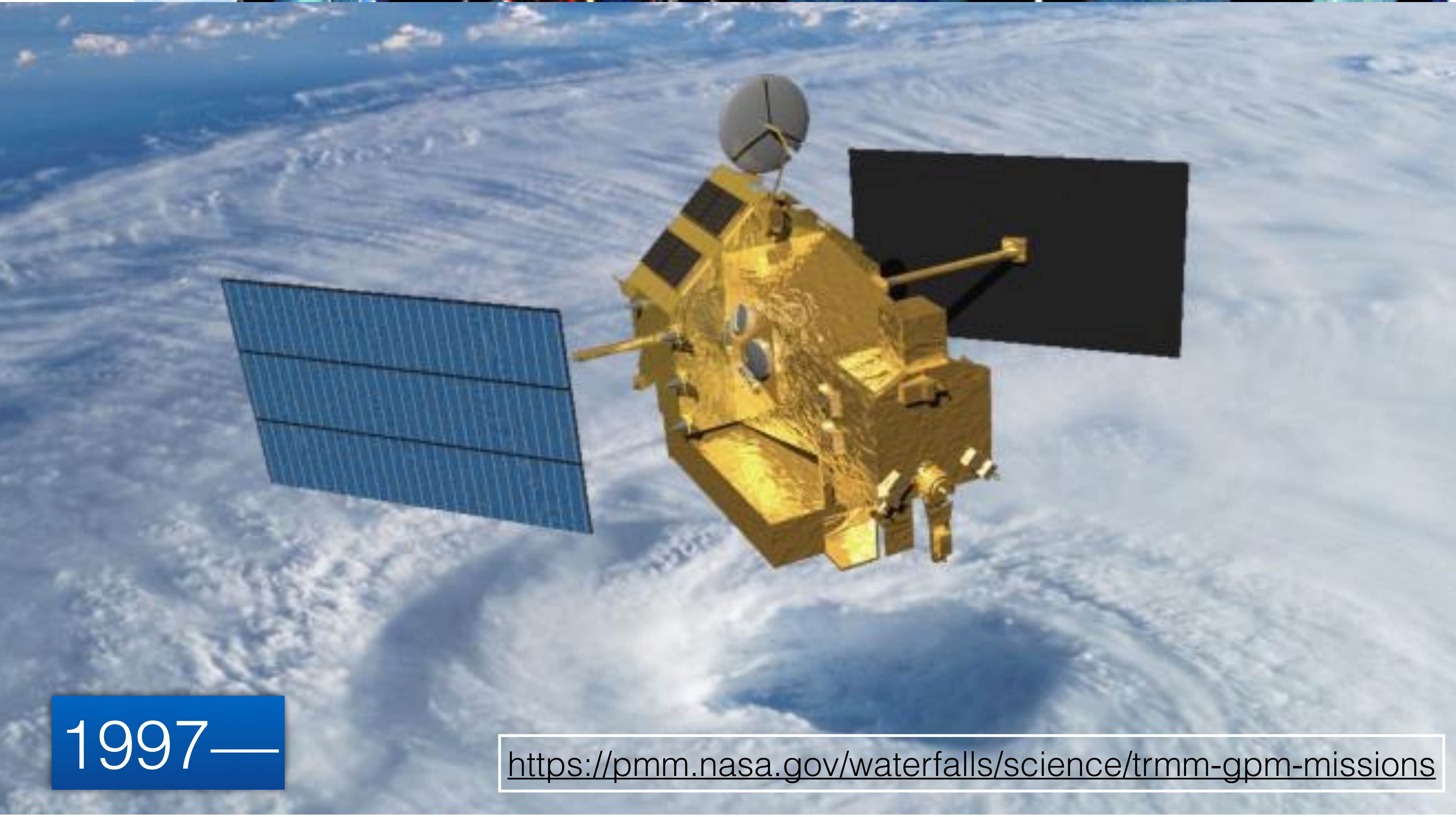
https://atrain.nasa.gov



2014—

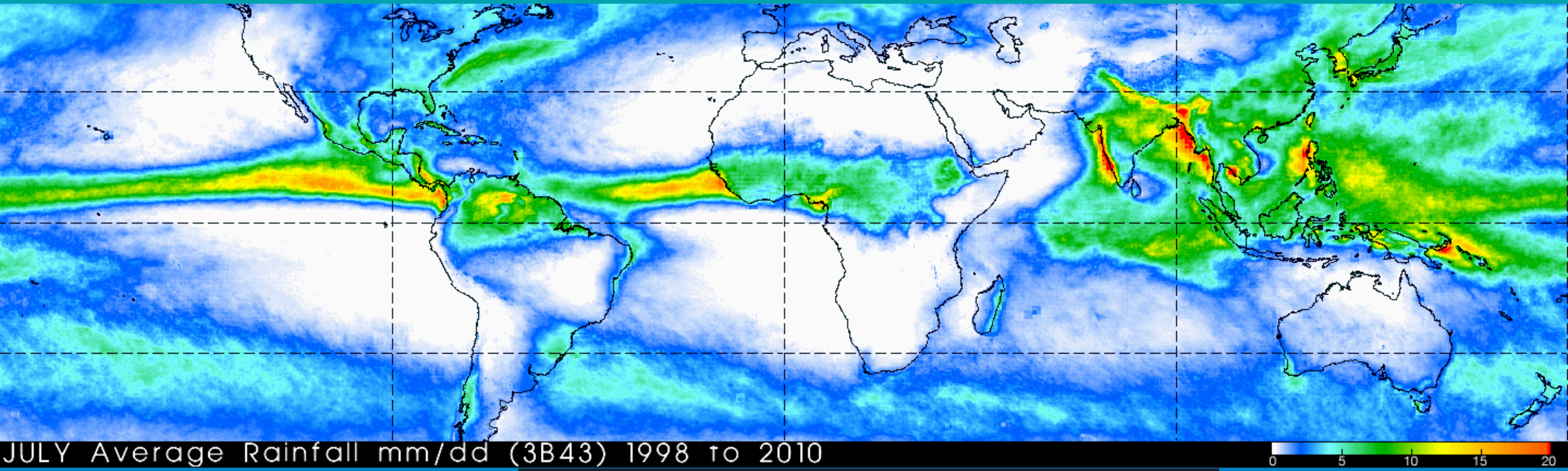
https://pmm.nasa.gov/waterfalls/science/trmm-gpm-missions



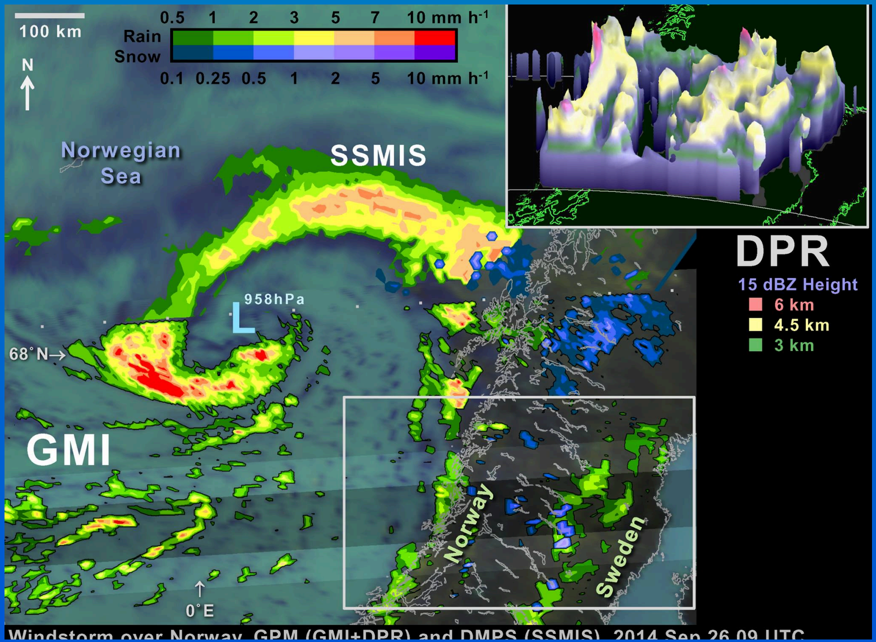


1997—

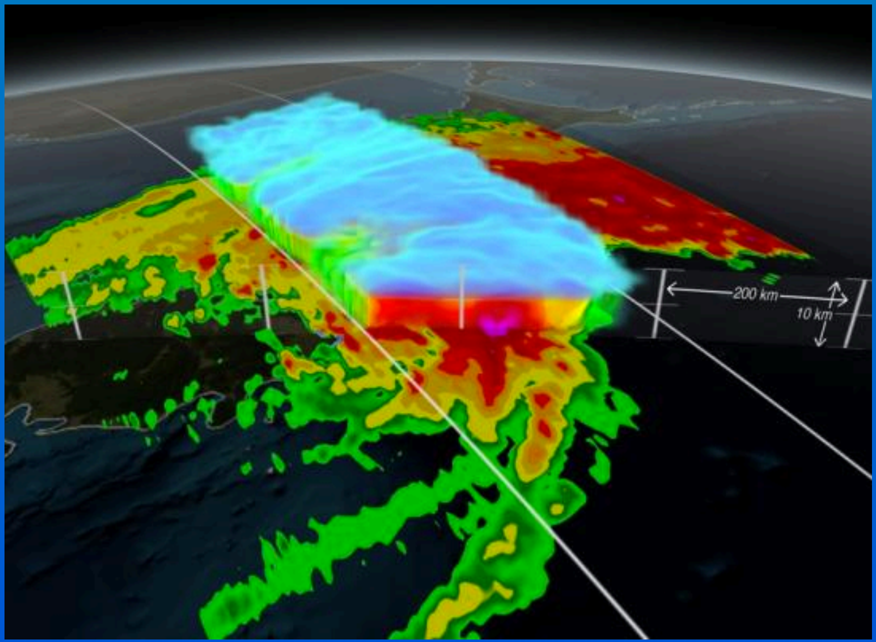
https://pmm.nasa.gov/waterfalls/science/trmm-gpm-missions



Climatological Precipitation in July (TRMM)



Extratropical Cyclone (2014, GPM)



Typhoon Phanfone (2014, GPM)

the data-science challenges

Multiple Spatial and Temporal Scales of Interests

aerosol- and cloud-radiative effects and forcings
cloud- and convection-coupled atmospheric motions
severe storms and extreme weather
climate change
extreme weather climatology
Intra-Seasonal Variability…

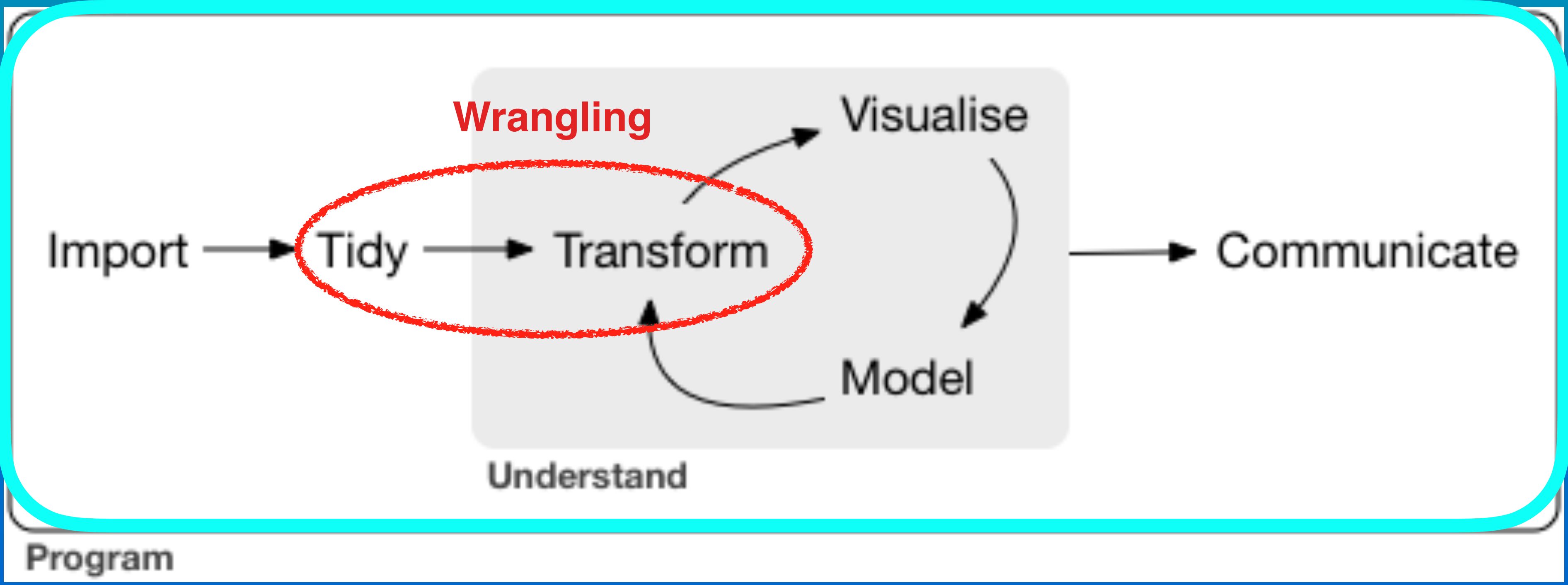
The global cloud and precipitation vary across a wide range of spatial and temporal scales, manifesting the complexity of the interacting processes within the water cycle.

In order to characterize it, we need:

- Global climatological records at spatial and temporal scales fine enough to resolve the local features of high-impact events

- Methods that allow deep analysis and detailed visualization of large complex data

Schematics of a typical data-science project



Grolemund and Wickham (2016) “R for Data Science”

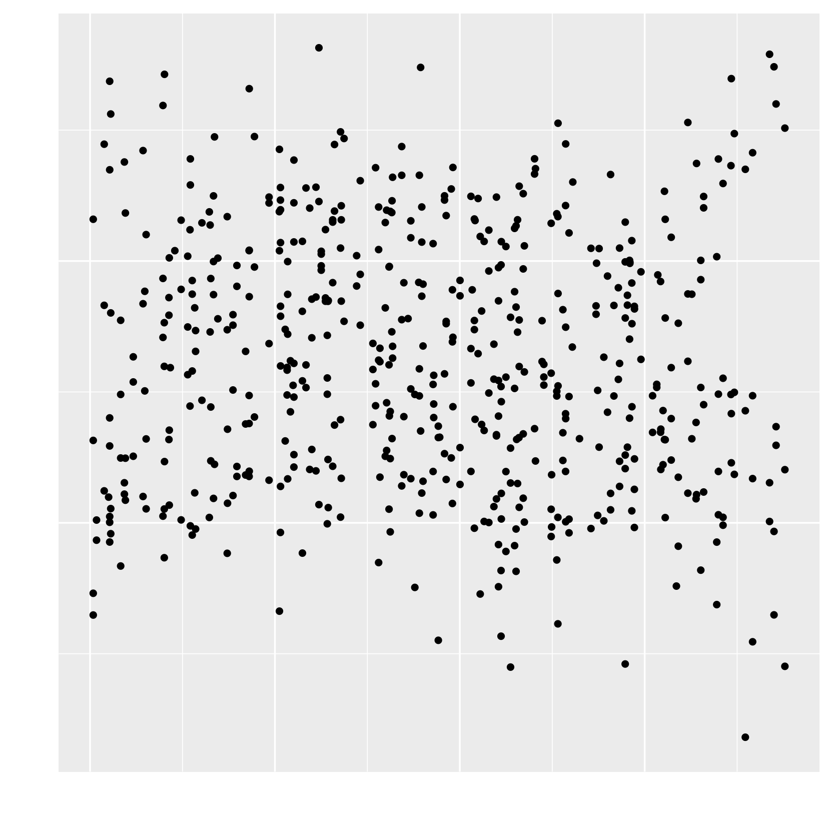
Analyze and Visualize Large Complex Data in R

δρ DeltaRho

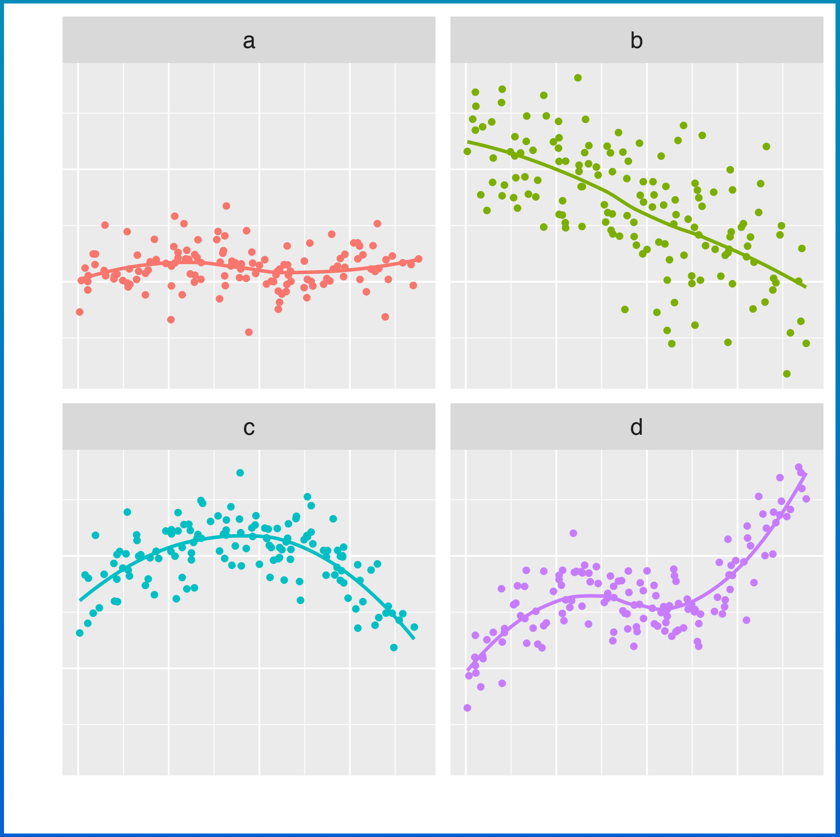
is an open source project to enable deep analysis and detailed visualization of large complex data in R.

http://deltarho.org

Division can reveal the structure in component parts of complex data

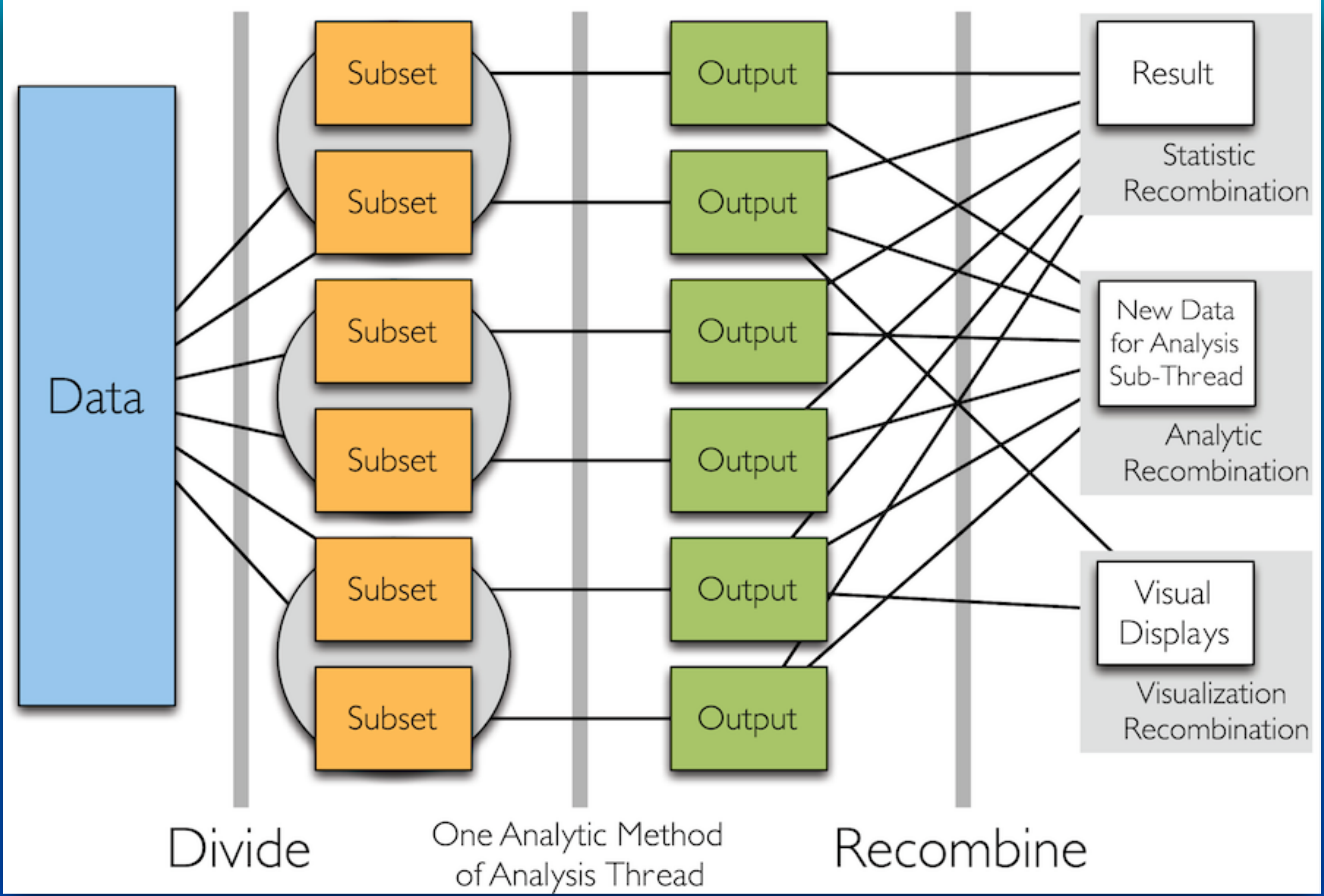


Full Data



Divided Data

DeltaRho is based on Divide and Recombine



hadoop

provides a scalable back end to power
the divide and recombine approach

- Hadoop distributed file system (HDFS)
- Parallel compute engine (Map/Reduce)

http://hadoop.apache.org

examples from small (100MB) to larger (250GB) data

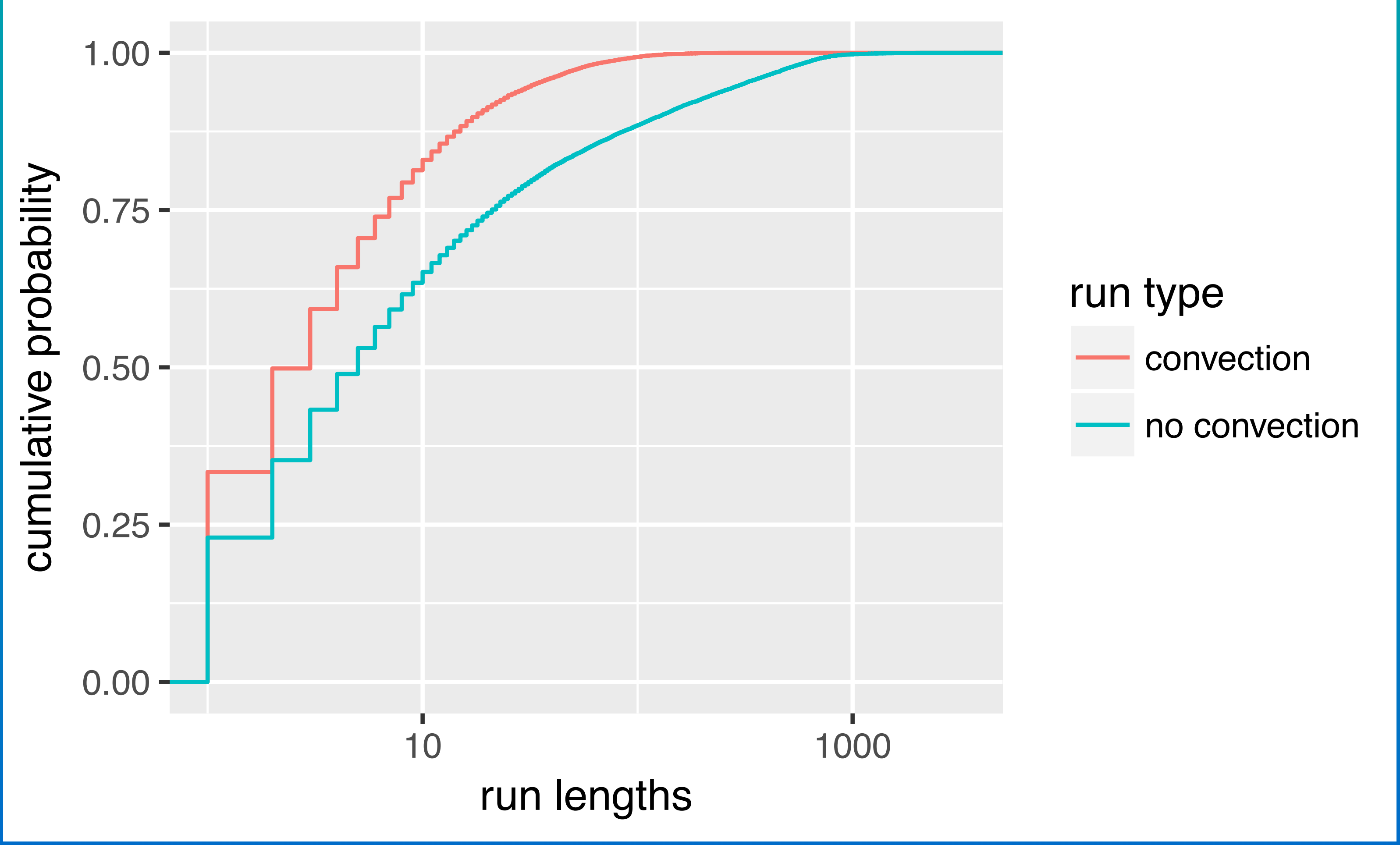
Data: Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations (CALIPSO)

- CALIPSO Standard 5km, 0.74s, Cloud Layer product, Level 2, V4.10
- CALIPSO Standard 5km, 0.74s, Aerosol Layer product, Level 2, V4.10
- Both subsets in the Northwestern Pacific domain and in 2015
- Each subset about ~150 MB

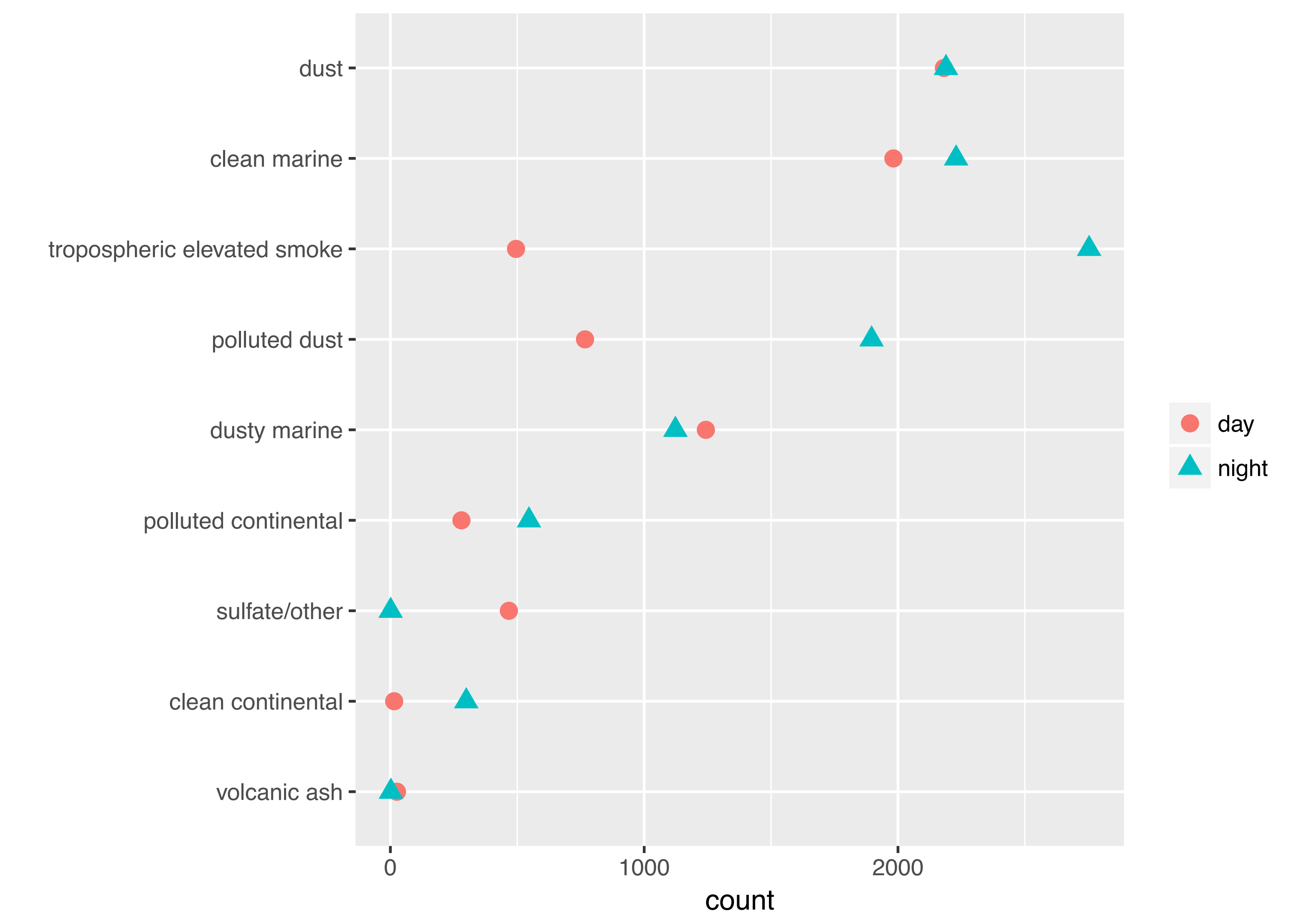
exploratory analysis: aerosol type and cloud type distributions

visualization: readily scaled up for detailed visualization

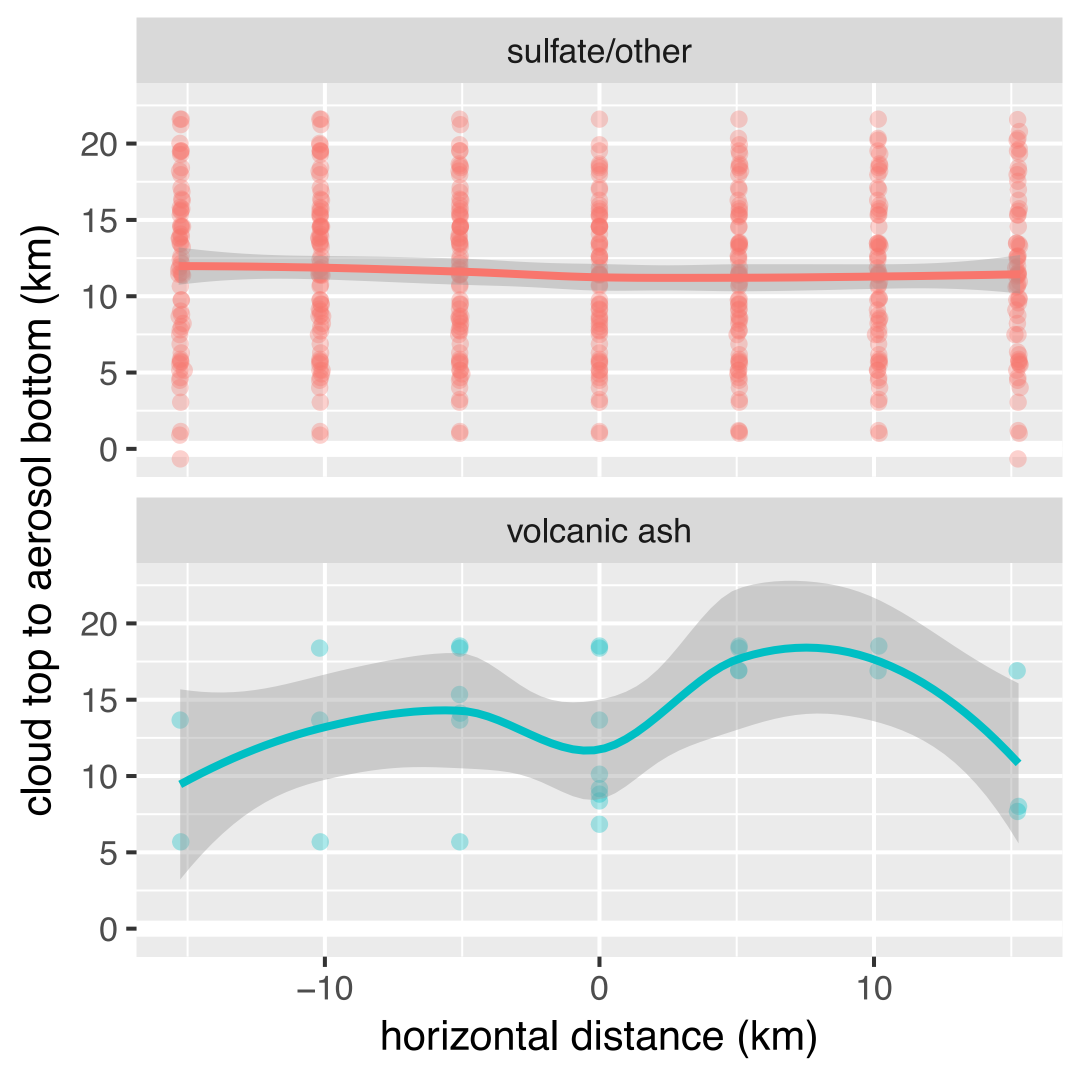
Humans: Matthew Bowers and Wanchen Wu
Machines: 1 Linux node/server with 8 cores (tidy)
1 MacPro with 6 cores (transform)
Program: R (datadr)



1/3 of Deep Convective shot runs are isolated, what are the aerosol type distributions within 30 kms around it?



Aerosol type distributions within 30 kms around isolated deep convection



Displacements of Stratospheric Aerosol Layers near an Isolated Deep-Convective Shot

Data: Tropical Rainfall Measuring Mission (TRMM)

- Version 7, 3B42, Multi-Satellite Precipitation Analysis (Huffman et al. 2007)

- Precipitation rate (mm/hr).

- 3-hourly 1998—2015

- 50° S—50° N, 180° W—180° E, 0.25° x 0.25° grid

- ~ 30 billion data points (250 GB)

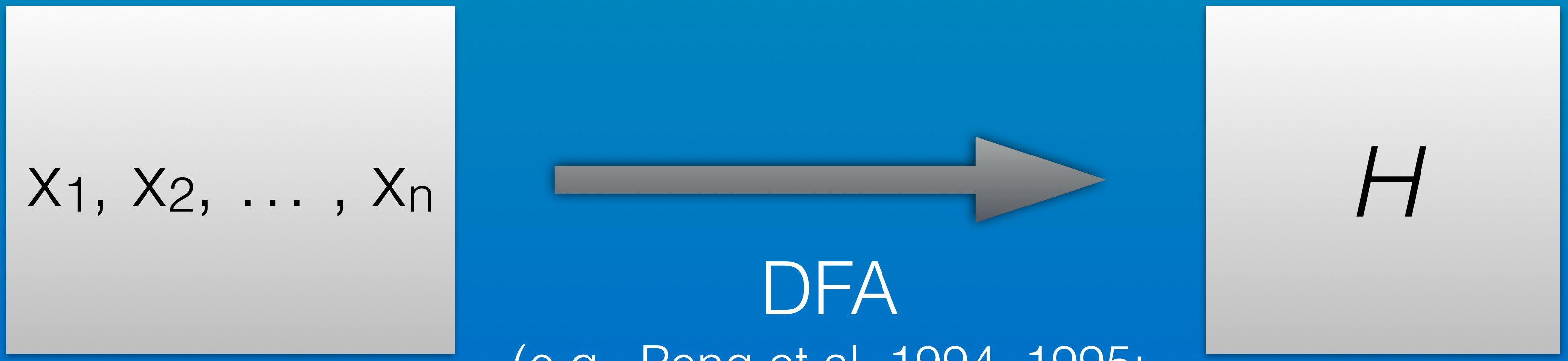
then, we asked:

What is the temporal correlation structure
of tropical precipitation?

How does it vary over the Earth?
in winter versus summer?
any longterm change over the years?

What does the longterm change mean?

An analytic method called Detrended Fluctuation Analysis (DFA) characterizes the temporal correlations



$X_1, X_2, \ldots, X_n$

H

DFA

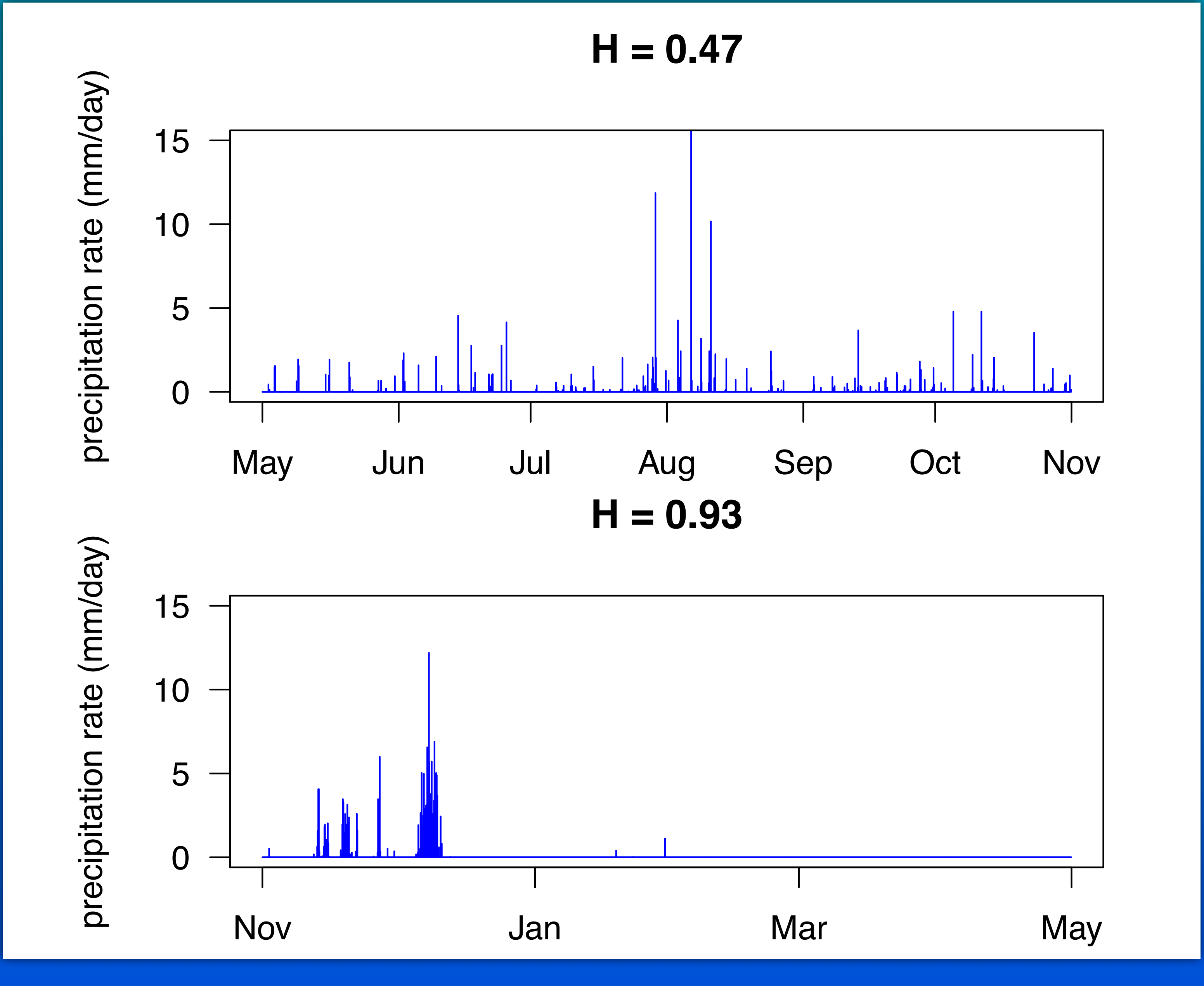
(e.g., Peng et al. 1994, 1995; Bowers et al. 2013)

Time Series Data

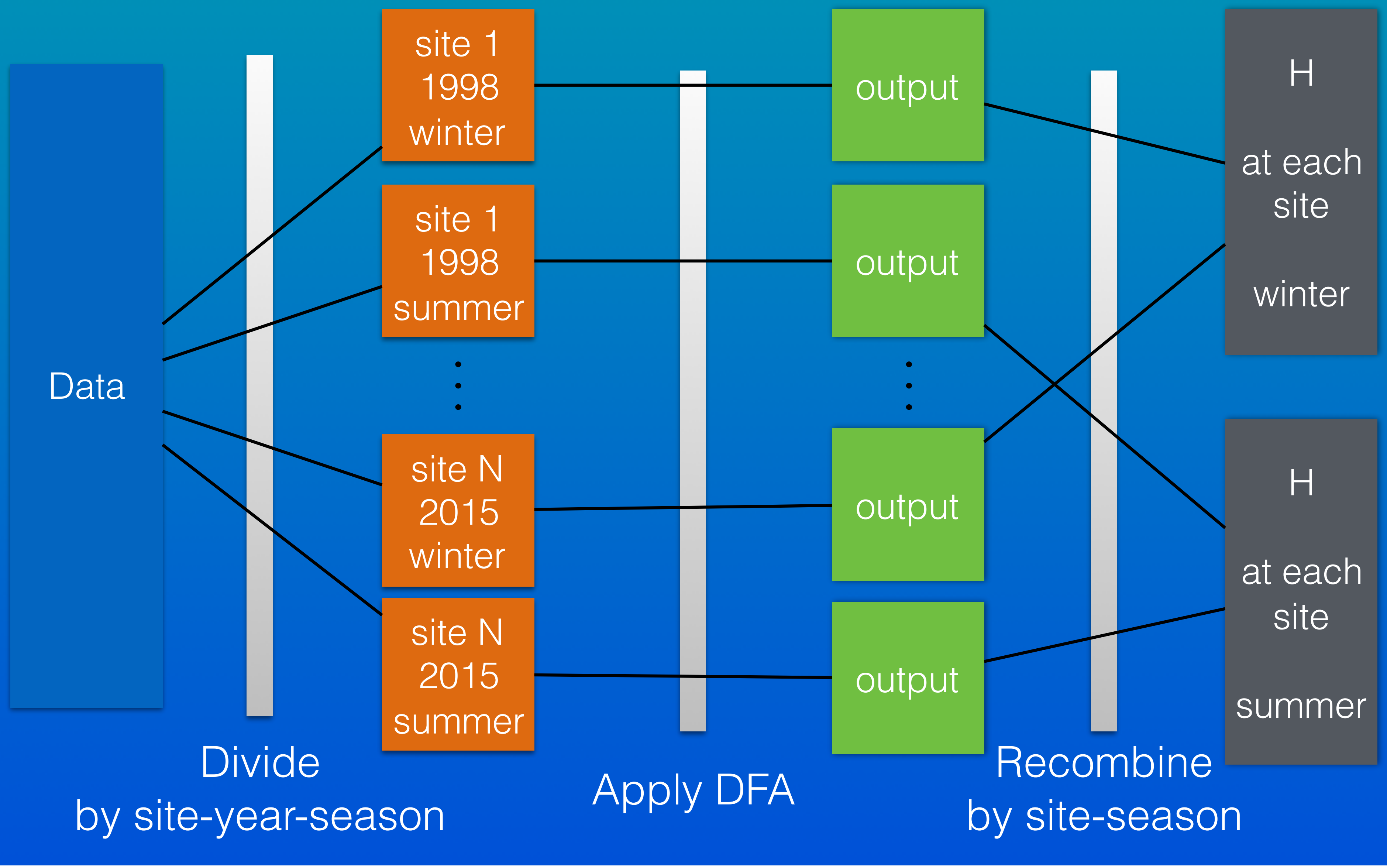
Hurst Parameter

DFA involves in detrending while varying time scales and a linear regression to find power-law scaling behavior characterized by the so-called Hurst parameter.

The value of Hurst Parameter indicates the degree of temporal clustering in precipitation.

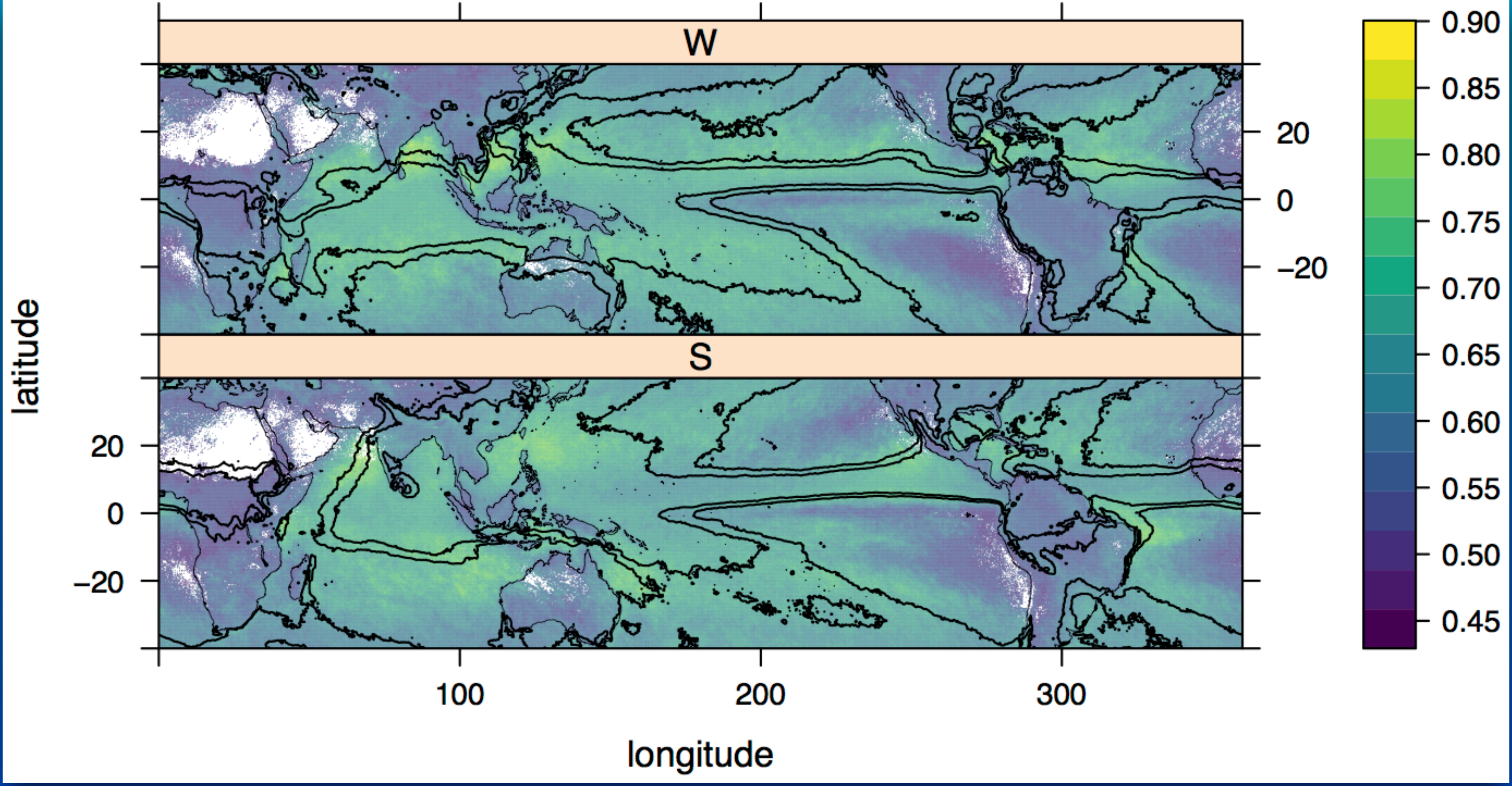


Divide by site-year-season, apply DFA, recombine statistically by site-season.

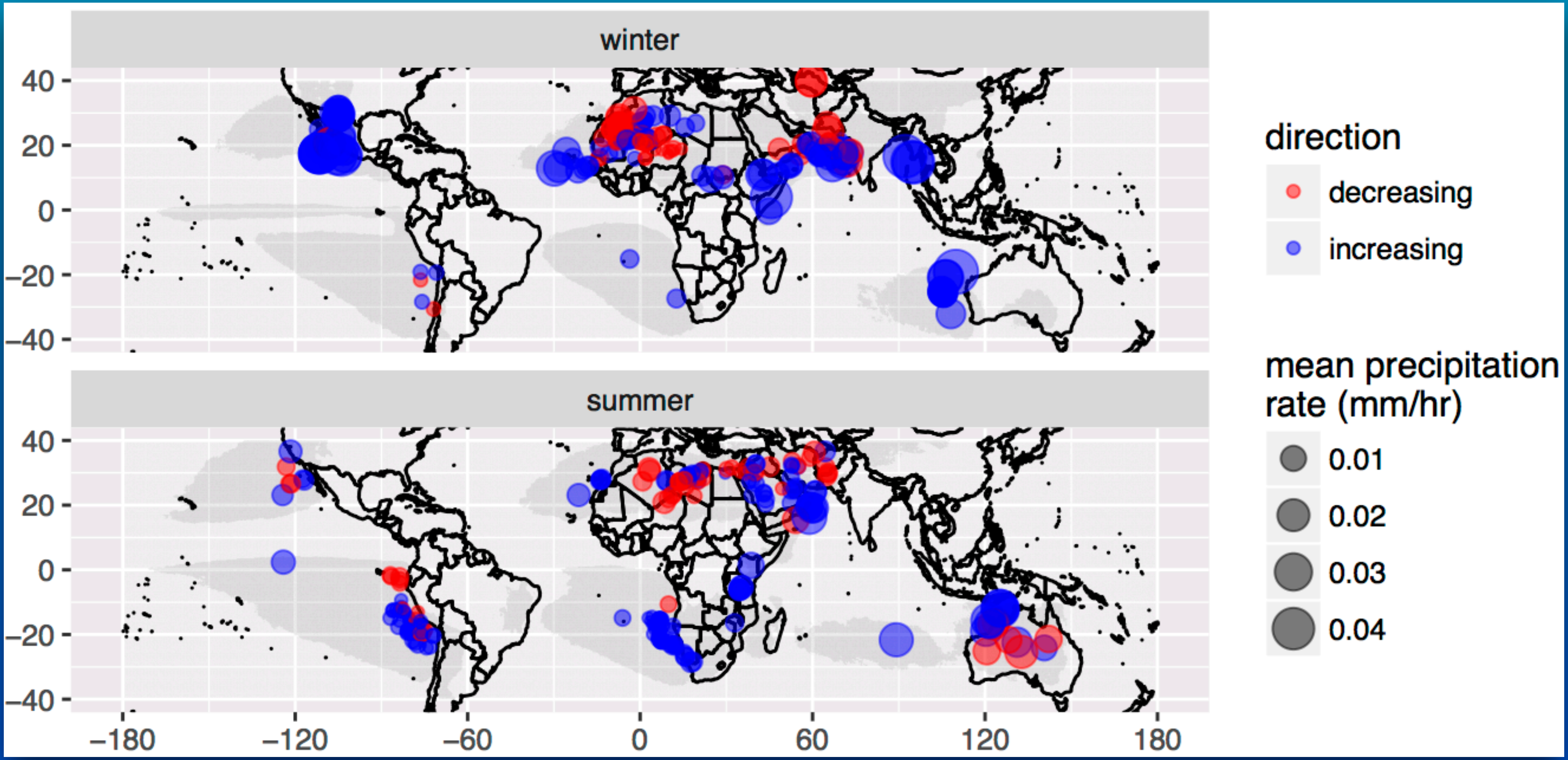


Regional features of H (for up to a month) emerge after statistical recombination

Seasonal Average H

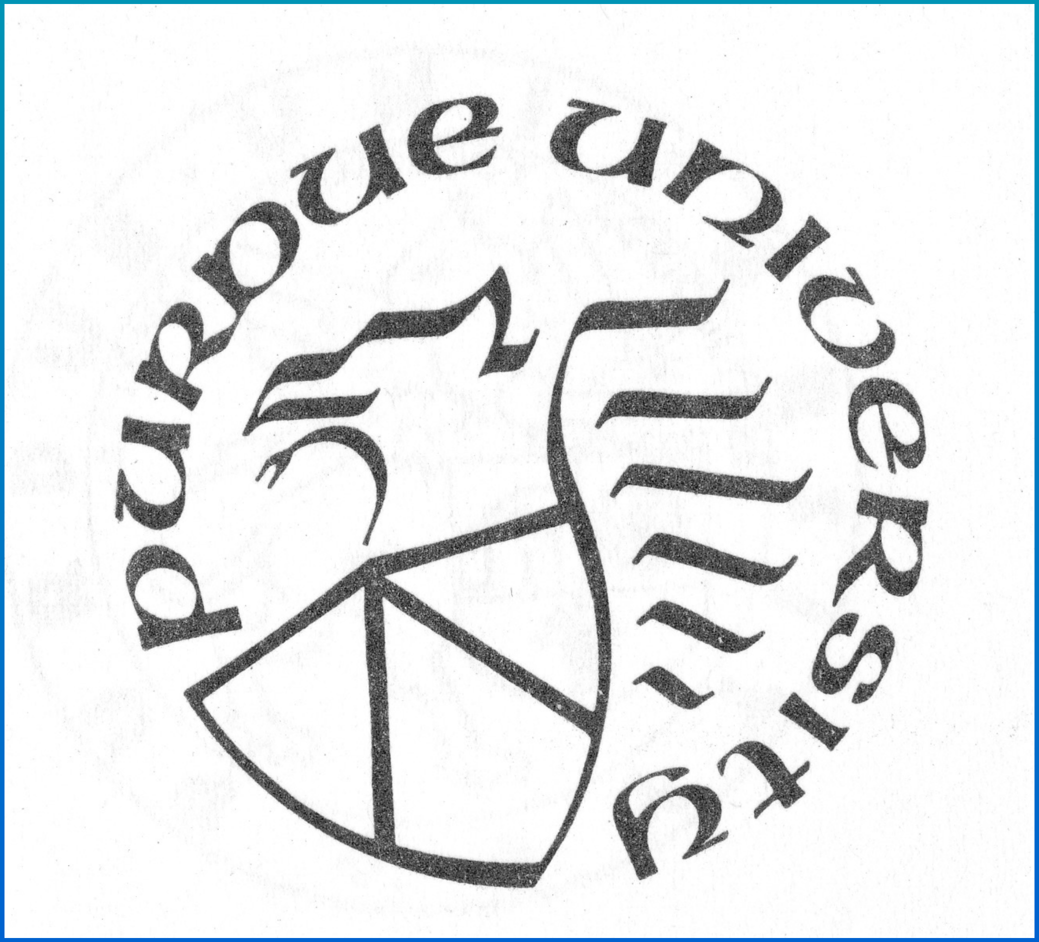


Locations with time change of H over 1998-2015 greater than 0.02 or less than -0.02 (per year)



Summary and Conclusions

- We are offering interdisciplinary data-science education and research opportunities to students utilizing geoscience data, especially at the graduate level
- At the undergraduate level, we plan to scaffold more geoscience students to be come ready to learn data science
- We teach Divide and Recombine to enable deep analysis of large complex data
    - Hadoop scales D&R to arbitrarily large datasets
- Graduate students in geosciences now can efficiently conduct amazing Big-data projects


PURDUE UNIVERSITY