Deep Analysis and Modeling of Satellite-based Precipitation — Opportunities for Short Term and Seasonal Rain Forecasts

Wen-wen Tung[1] and William S. Cleveland[2,3]

[1]Department of Earth, Atmospheric, & Planetary Sciences,
[2]Department of Statistics,
[3]Department of Computer Science,
Purdue University, West Lafayette, IN

Acknowledgment: Matthew C. Bowers

About this talk

- Bill
  - Deep analysis with DeltaRho (http://deltarho.org)
  - Introduction to data-driven stochastic/statistical modeling using rainfall as an example
- Wen-wen
  - Stochastic parameterization in weather and climate models
  - Deep analysis of TRMM data using DeltaRho

A few obvious applications of data-driven statistical modeling study

- Model Output Statistics (MOS)— is a type of statistical post-processing techniques used to improve numerical weather models' ability to forecast by relating model outputs to observational or additional model data (Glahn and Lowry, 1972)

- Downscaling — dynamical or statistical procedures to take information known at large scales to make predictions at local (fine) scales. Statistical downscaling consists of i) the development of statistical relationships between local variables (e.g., surface air temperature and precipitation) and large-scale predictors (e.g., pressure fields), and ii) the application of such relationships to the output of global/coarser-resolution climate or NWP models to estimate local/fine-scale characteristics

- Stochastic parameterizations — having been used in several numerical weather prediction models, including some run operationally, and found to improve forecast skill by both increasing the spread of ensemble forecasts and reducing the size of errors of the ensemble mean forecasts (e.g., Buizza et al., 1999; Palmer et al., 2009; Reynolds et al., 2011; YoneharaandM. Ujiie, 2011; Bouttier et al., 2012; Sušelj et al., 2014; Berner et al., 2015; Sanchez et al., 2016).

Adding stochasticity to atmospheric parameterization schemes improves simulated tropical climate (Rev. Berner et al. 2017)

- Berner et al. (2008) and Weisheimer et al. (2014)—stochastic parameterizations applied in the ECMWF coupled atmosphere-ocean model reduce biases in tropical mean rainfall
- Sanchez et al. (2016) found a similar result in the Met Office atmospheric model.
- Lin and Neelin (2000, 2002)—including stochasticity in the convection parameterization of intermediate complexity general circulation models (GCMs) improves aspects of the tropical variability; Lin and Neelin (2003) also showed this in NCAR community climate model.
- Davini et al. (2017)—stochastic parameterizations in EC-Earth improve the simulation of tropical rainfall rate distributions and the Madden-Julian Oscillation
- Wang et al. (2016)— the scheme of Plant and Craig (2008) improves the simulated tropical rainfall rate distribution in the NCAR CAM
- Dorrestijn et al. (2016), Goswami et al. (2016), and Peters et al. (2017) showed that variants of the stochastic multicloud model of Khouider et al. (2010) improved aspects of tropical variability simulated in different GCMs, and Frenkel et al. (2012) showed similar results in a single-column model context.
- Christensen et al. (2017)—stochastic physics greatly reduced excessive El Niño variability in the NCAR CAM4

Common stochastic physics schemes

- Stochastically perturbed parameterization tendencies scheme (SPPT) (Buizza et al., 1999; Palmer et al., 2009)

- Stochastic kinetic energy backscatter scheme (SKEBS) (Shutts, 2005; Berner et al., 2009)

- Both are additions to existing closure assumptions in convection parameterization

Stochastically perturbed parameterization tendencies scheme (SPPT)

- Treats the total parameterized tendencies in prognostic model variables, such as temperature, humidity, and wind that are produced by the subgrid parameterizations as uncertain quantities.
- The tendencies are multiplied by a random number that scales them up or down.
    - The random number is typically correlated in space and time, to account for the fact that parameterization errors are spatially and temporally correlated.
- Rainfall is not directly perturbed, but perturbations to atmospheric tendencies will affect rainfall at subsequent time steps.

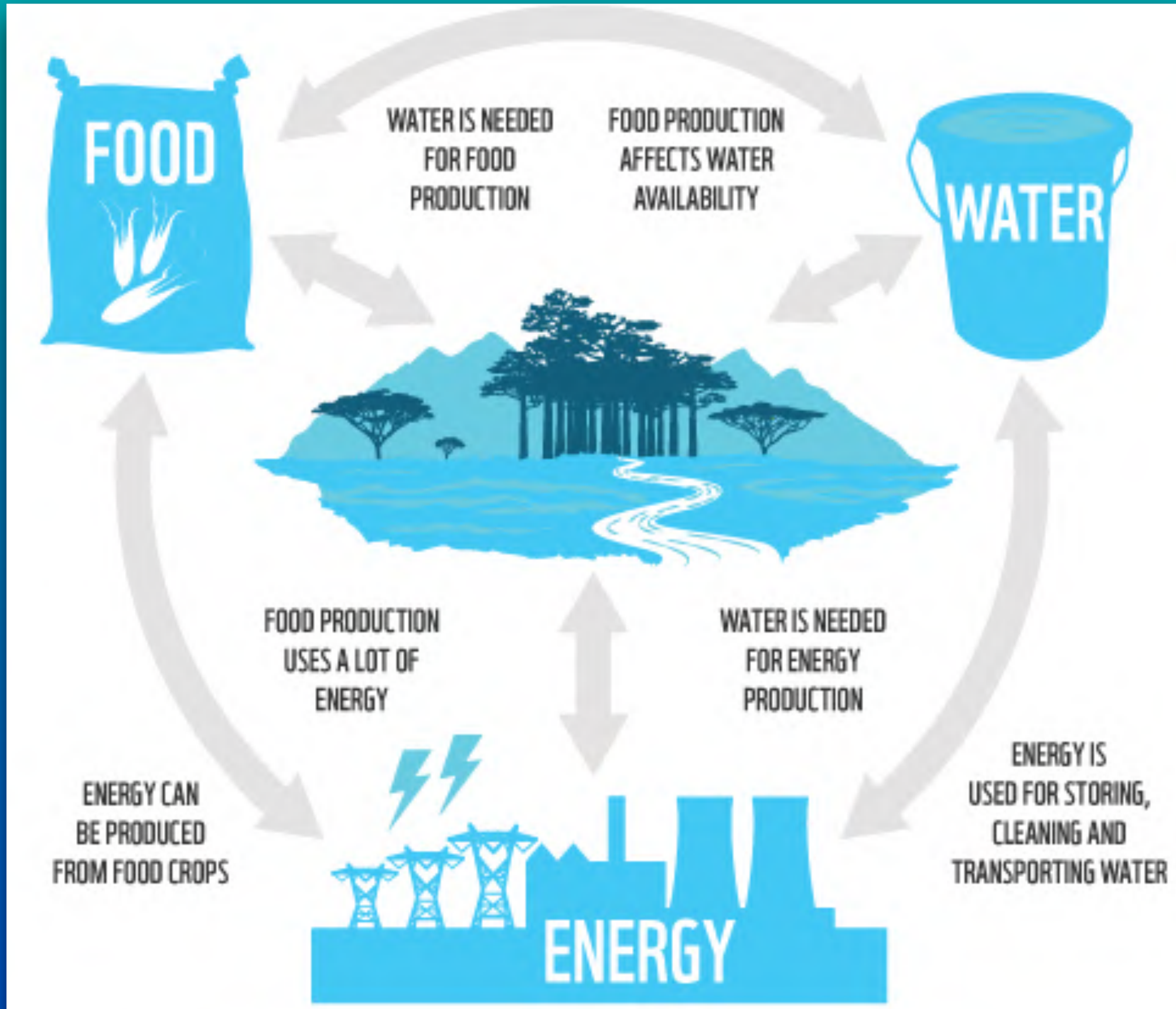
Stochastic kinetic energy backscatter scheme (SKEBS)

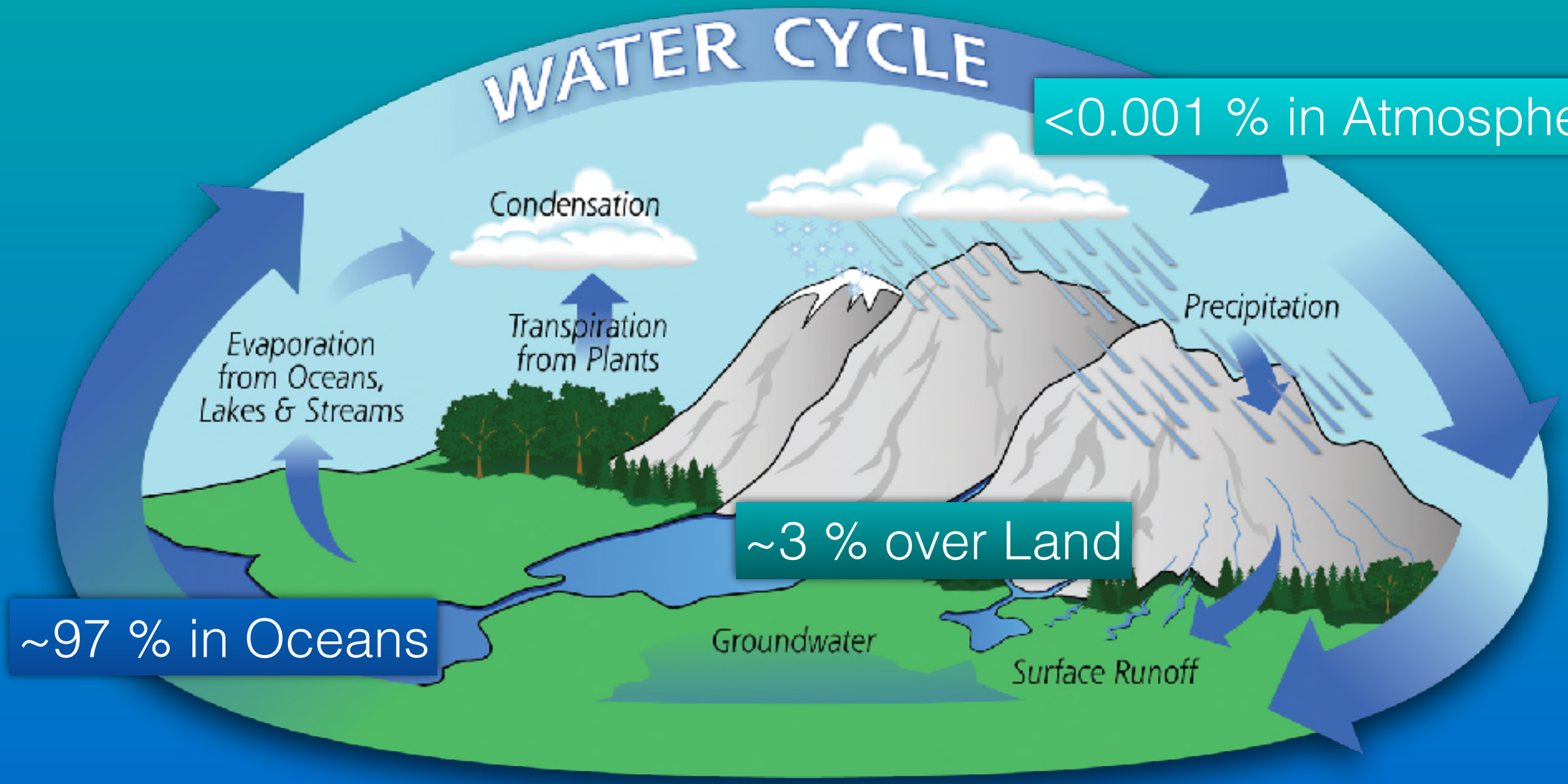
- Provides a representation of errors resulting from energy dissipation by subgrid-scale processes affecting larger scales—a process that is not parameterized in deterministic atmospheric models.
- The atmospheric stream function tendency is perturbed at each grid point and time step by a random pattern with a specified amplitude, spatial power spectrum, and temporal autocorrelation. (The temperature may also be perturbed, depending on the implementation.)
- The amplitude of the perturbations may depend on dissipative processes acting at each location associated with the model numerics and subgrid-scale processes such as convection.
- These perturbations to the model state will influence the behavior of subgrid parameterization schemes and hence affect precipitation.

“More reliable forecasts with less precise computations: a fast-track route to cloud-resolved weather and climate simulators?” — Owing to the lack of scale separation in atmospheric energy on scales of hundreds of kilometers and less, the closure schemes for weather and climate simulators should be based on stochastic–dynamic systems rather than deterministic formulae.

T. N. Palmer (2014)

Motivation for us to study precipitation

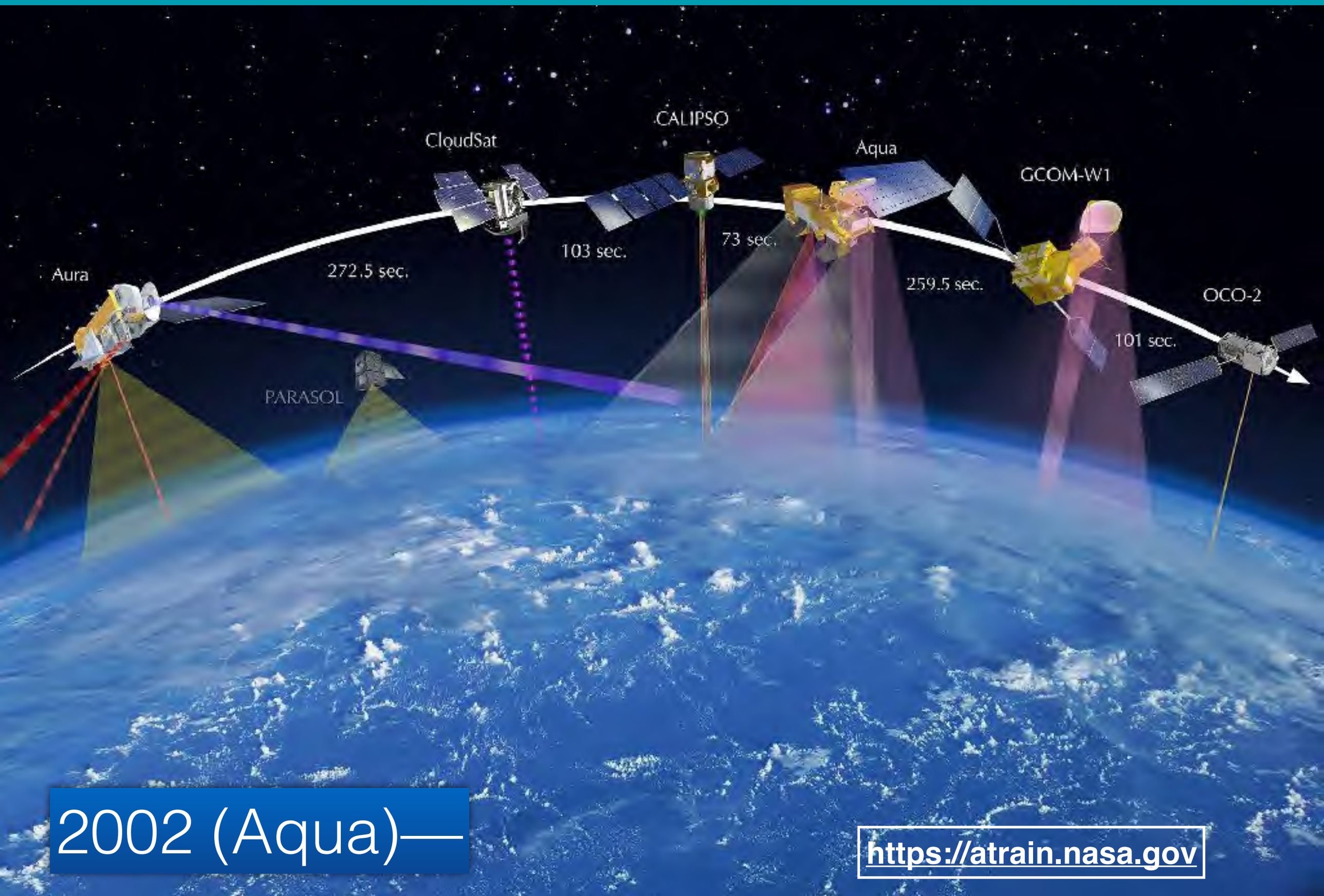

FOOD
WATER IS NEEDED FOR FOOD PRODUCTION
FOOD PRODUCTION AFFECTS WATER AVAILABILITY
WATER
FOOD PRODUCTION USES A LOT OF ENERGY
WATER IS NEEDED FOR ENERGY PRODUCTION
ENERGY CAN BE PRODUCED FROM FOOD CROPS
ENERGY IS USED FOR STORING, CLEANING AND TRANSPORTING WATER
ENERGY



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

The Afternoon Constellation — A-Train



2002 (Aqua)—

https://atrain.nasa.gov

全球降水観測計画

GPM

GLOBAL PRECIPITATION MEASUREMENT

2014—

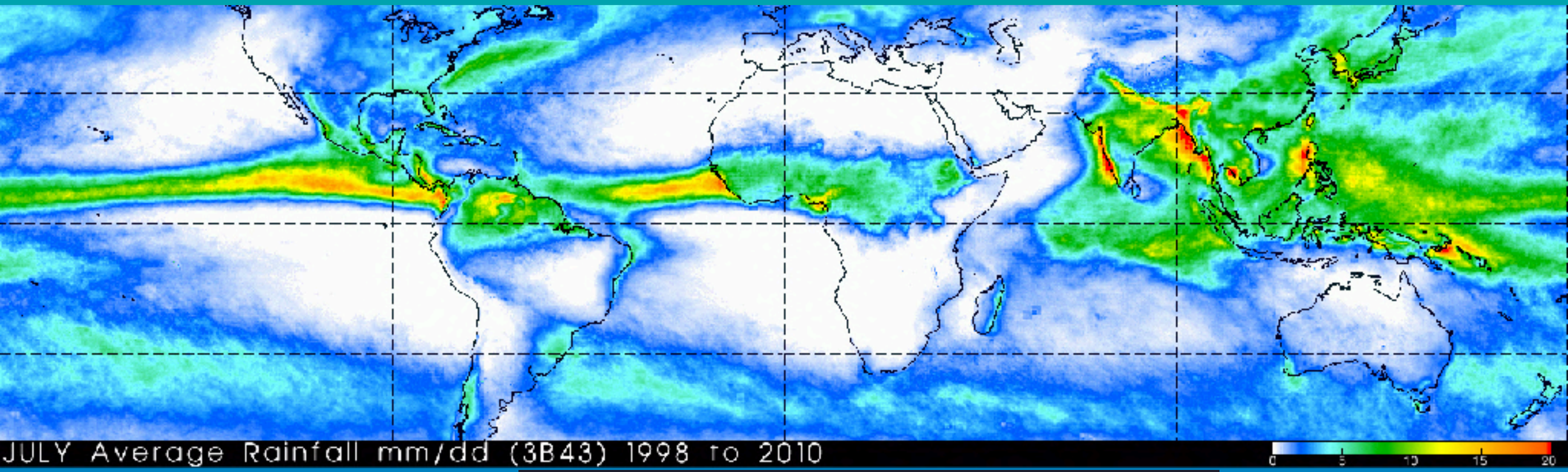
https://pmm.nasa.gov/waterfalls/science/trmm-gpm-missions

TRMM

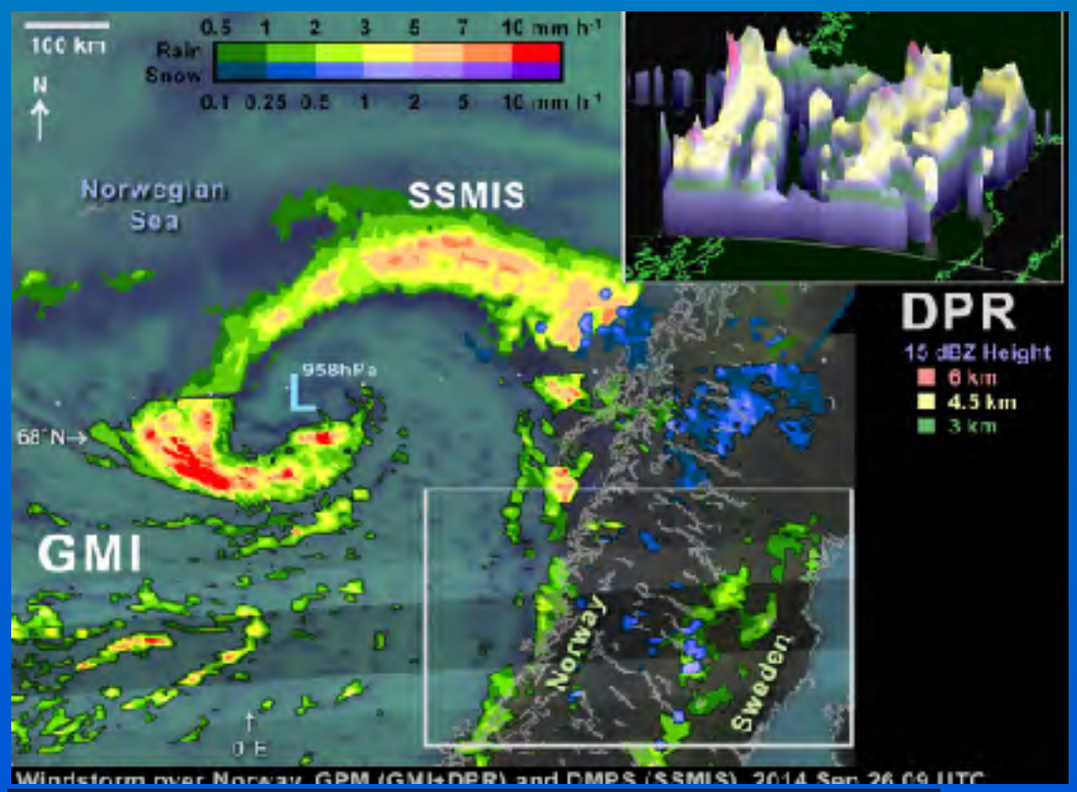
Tropical Rainfall Measuring Mission

1997—

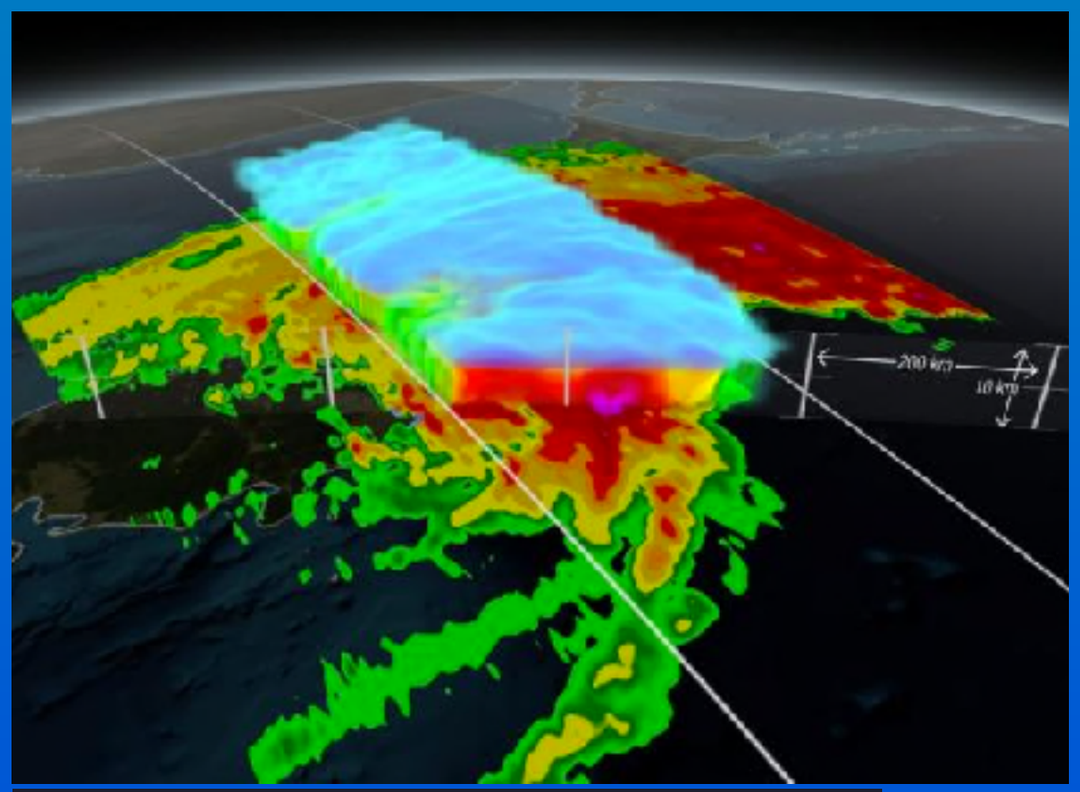
https://pmm.nasa.gov/waterfalls/science/trmm-gpm-missions



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
Intra-Seasonal Variability
Subseasonal to Seasonal (S2S) Forecasts
climate change/
extreme weather climatology

A large portion of S2S predictability originates from:

- Natural modes of variability (e.g., El Nino-Southern Oscillation [ENSO], the Madden-Julian Oscillation [MJO], and the Quasi Biennial Oscillation [QBO])

- Slowly-varying processes (e.g., involving soil moisture, snow pack and other aspects of the land surface, ocean heat content, currents and eddy positions, and sea ice)

- Elements of external forcing (e.g., aerosols, greenhouse gasses) that can result in a systematic and predictable evolution of the Earth system

- Most interacting with precipitation and precipitation process, and with potential for stochastic modeling/parameterization

  Next Generation Earth System Prediction: Strategies for Subseasonal to Seasonal Forecasts (2016 NRC report, http://www.nap.edu/21873)

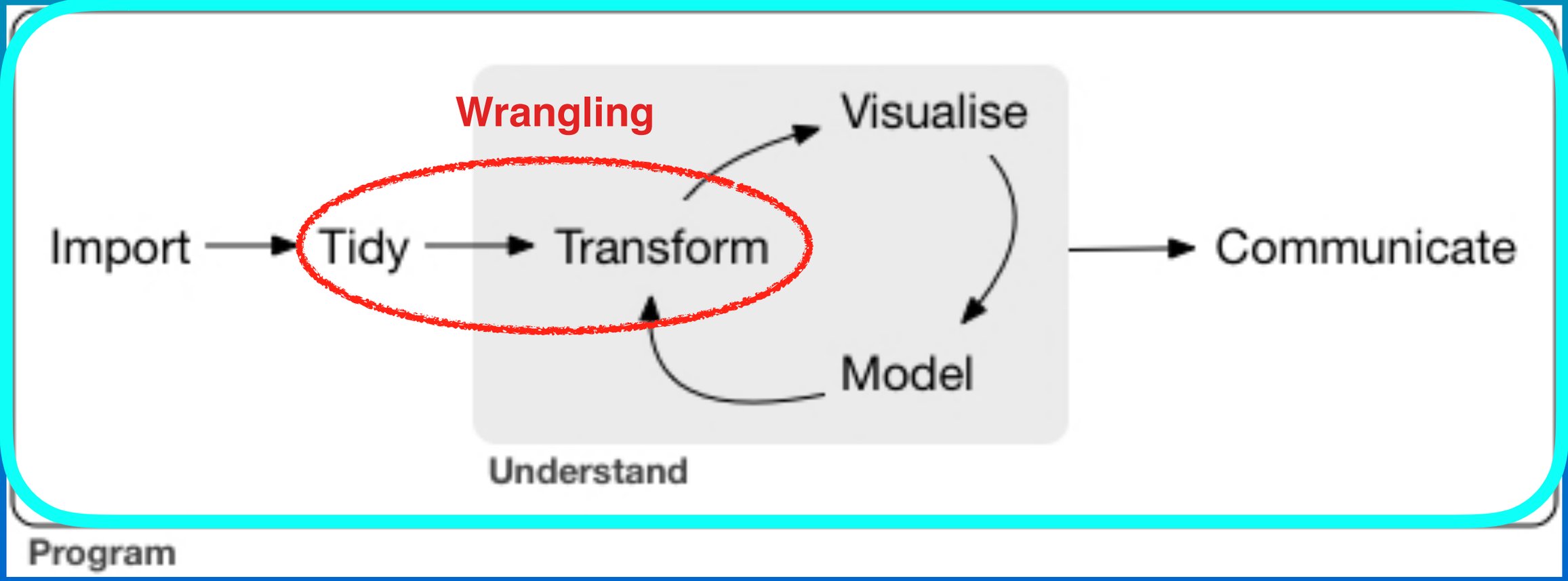
The global precipitation vary across a wide range of spatial and temporal scales, manifesting the complexity of the interacting processes within the water cycle.

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”

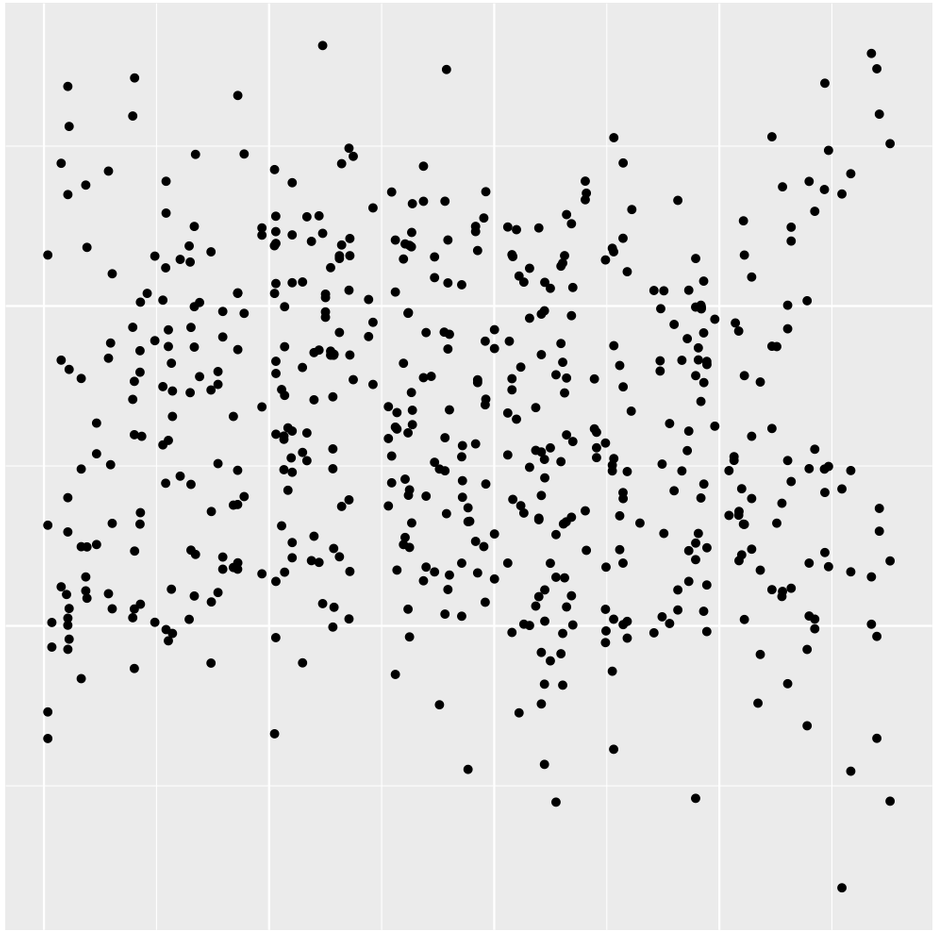
Analyze and Visualize Large Complex Data in R

δρ DeltaRho

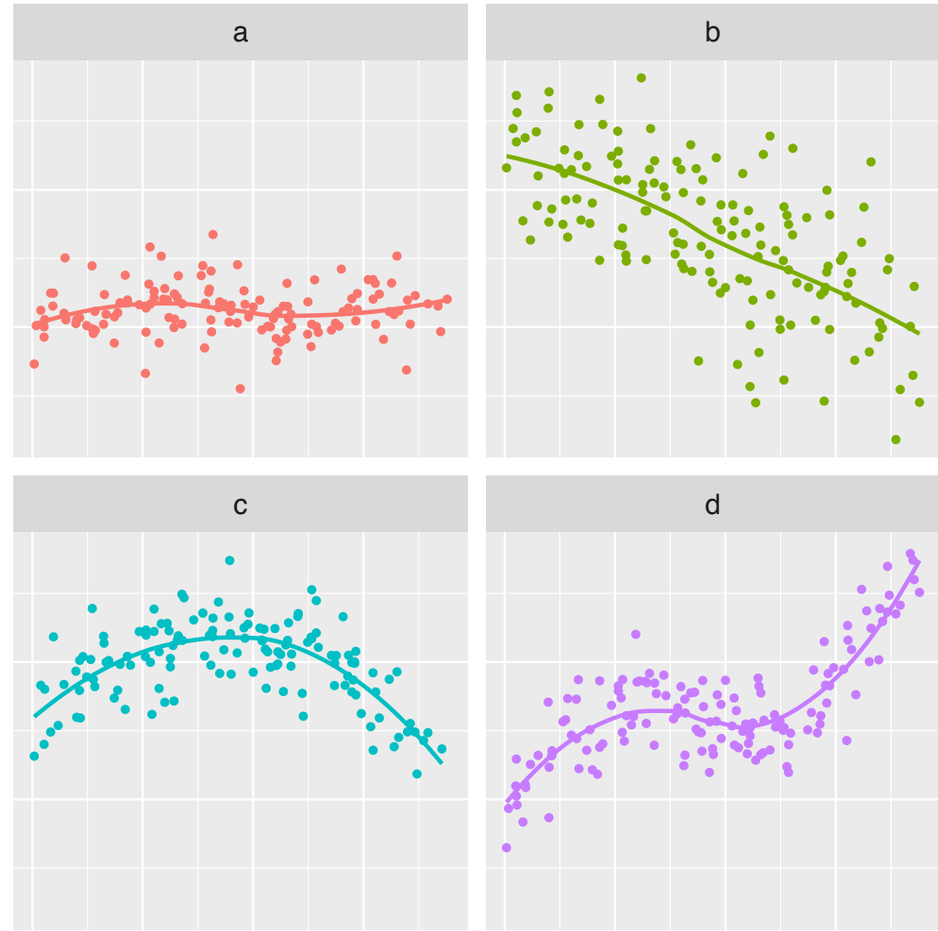
is an open source project to enable deep analysis and detailed visualization of large complex data in R.

http://deltarho.org

Division can reveal the structure in component parts of complex data

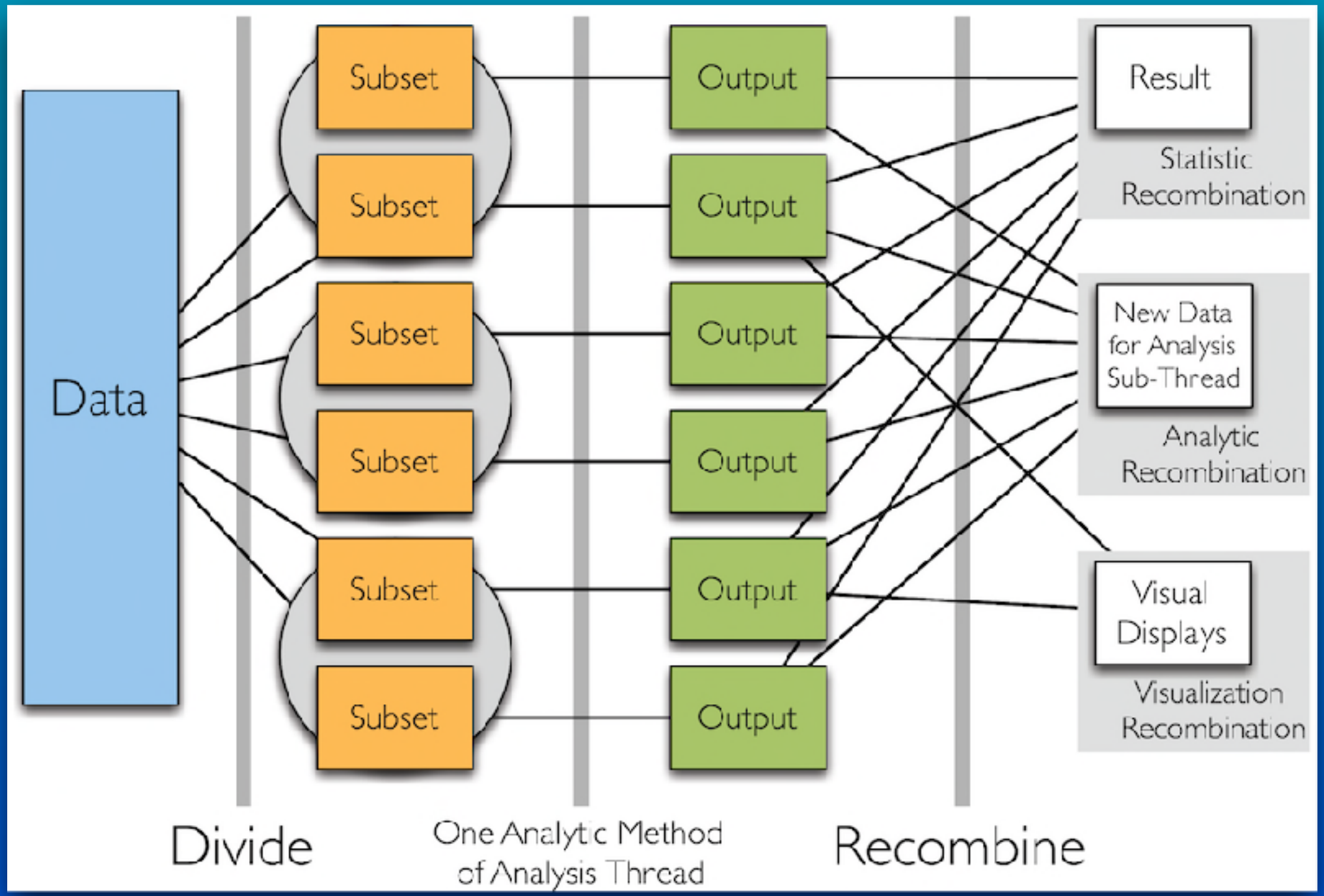


Full Data



Divided Data

DeltaRho is based on Divide and Recombine



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

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)

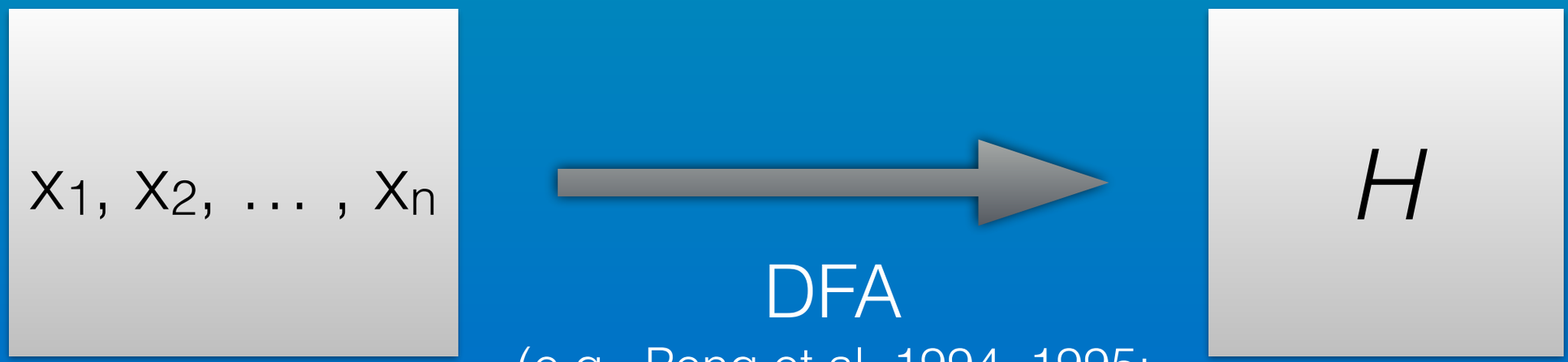
then, we asked:

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

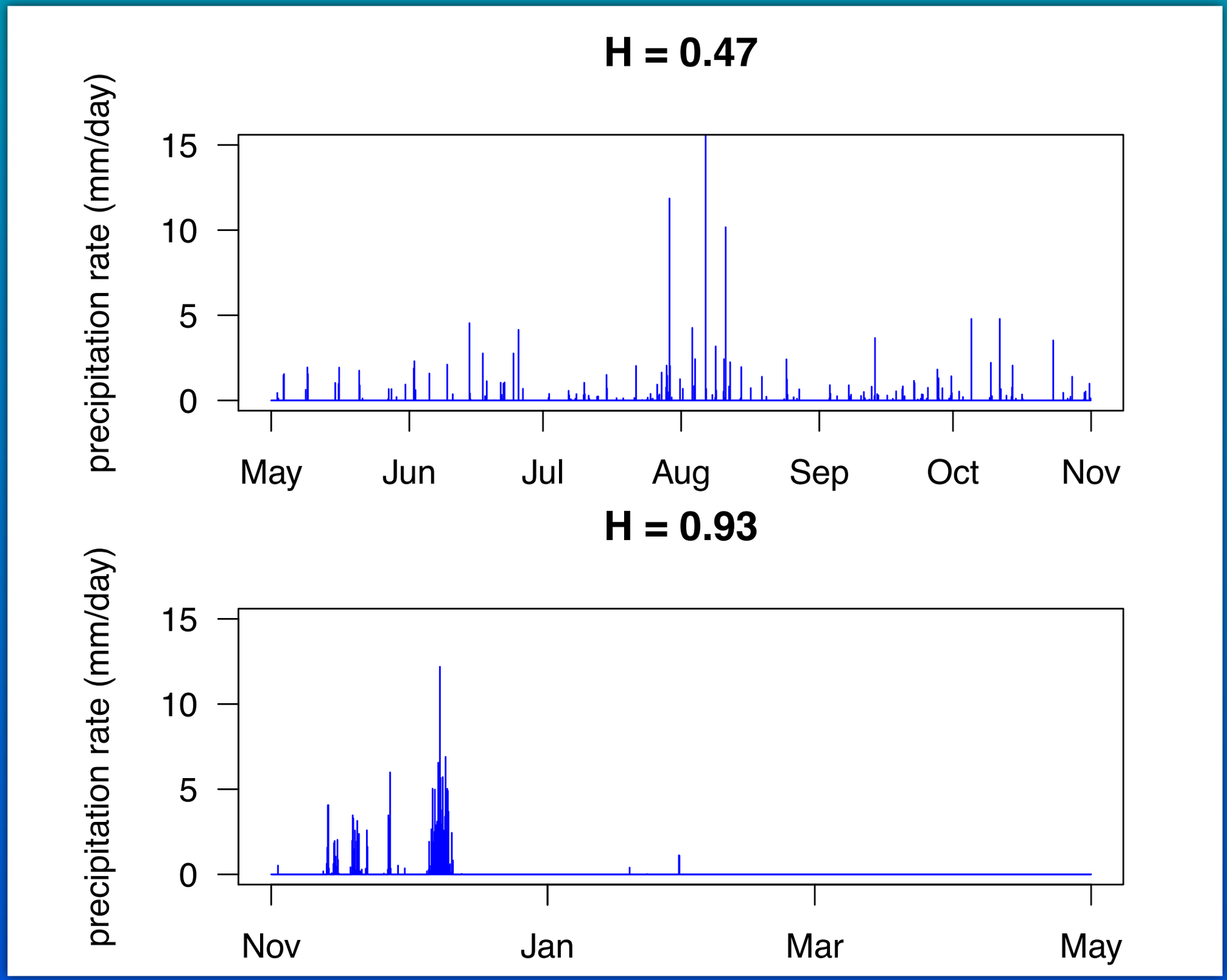


Time Series Data

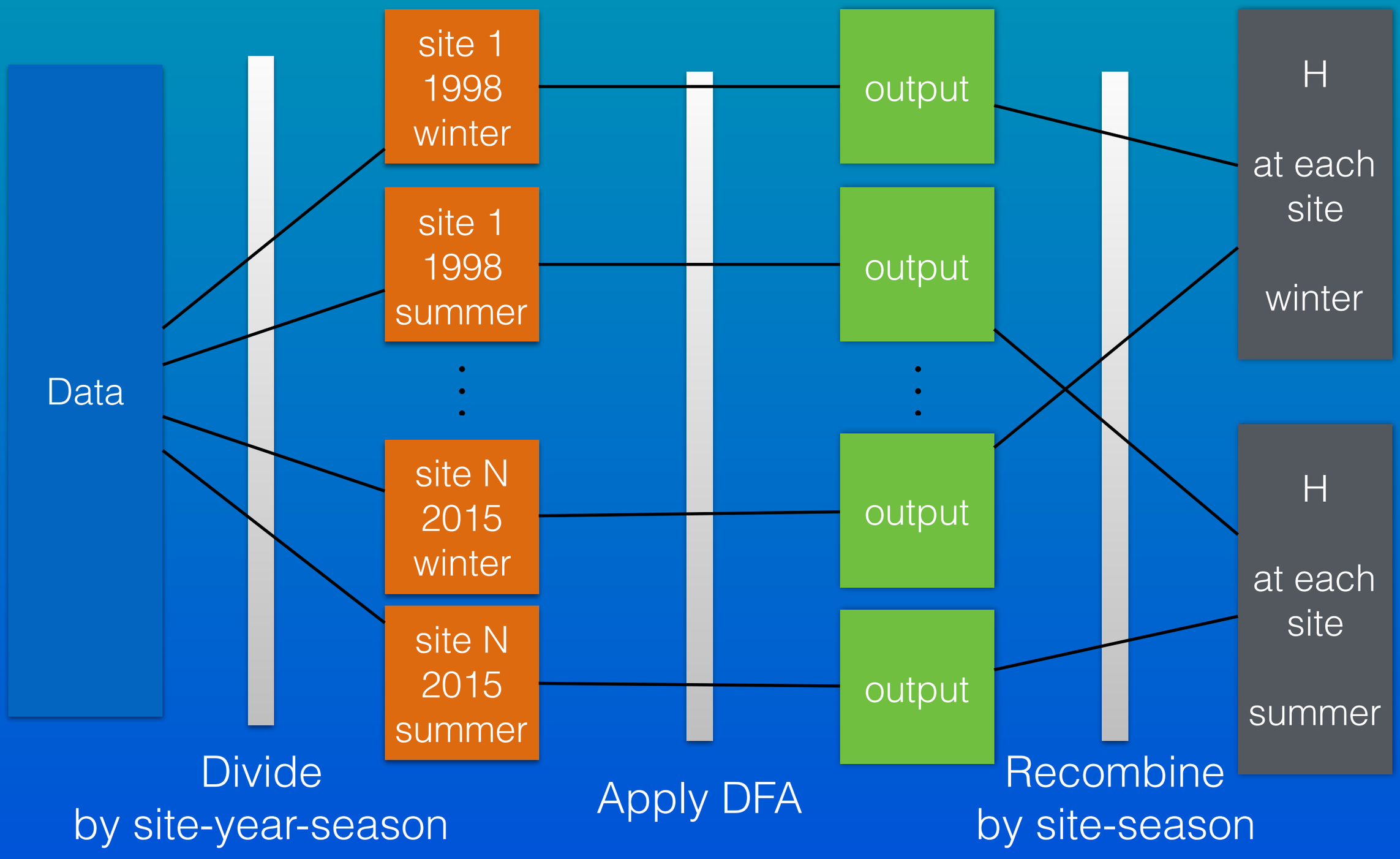
Hurst Parameter

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.

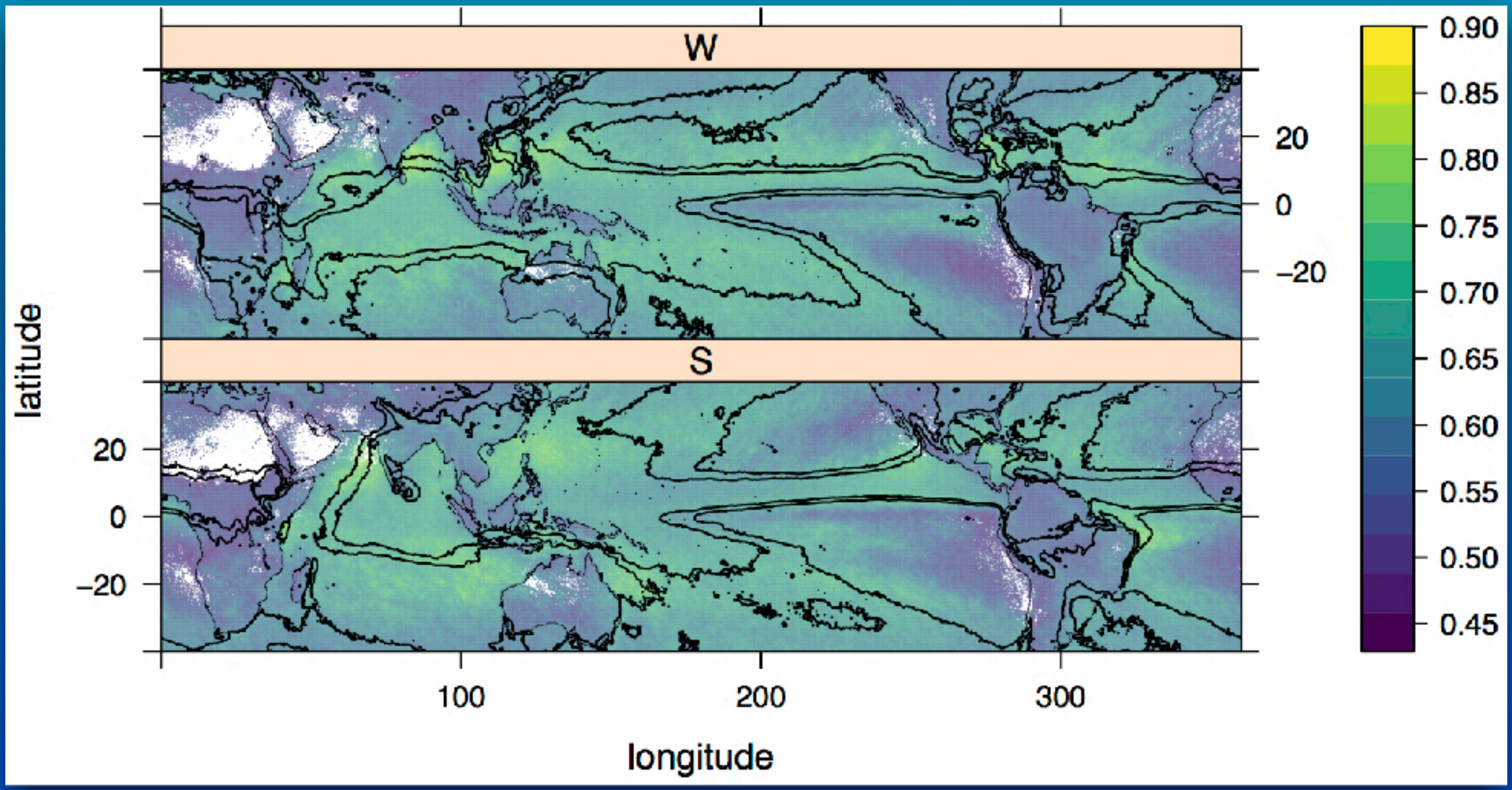
The value of Hurst Parameter indicates the degree of temporal clustering in precipitation.



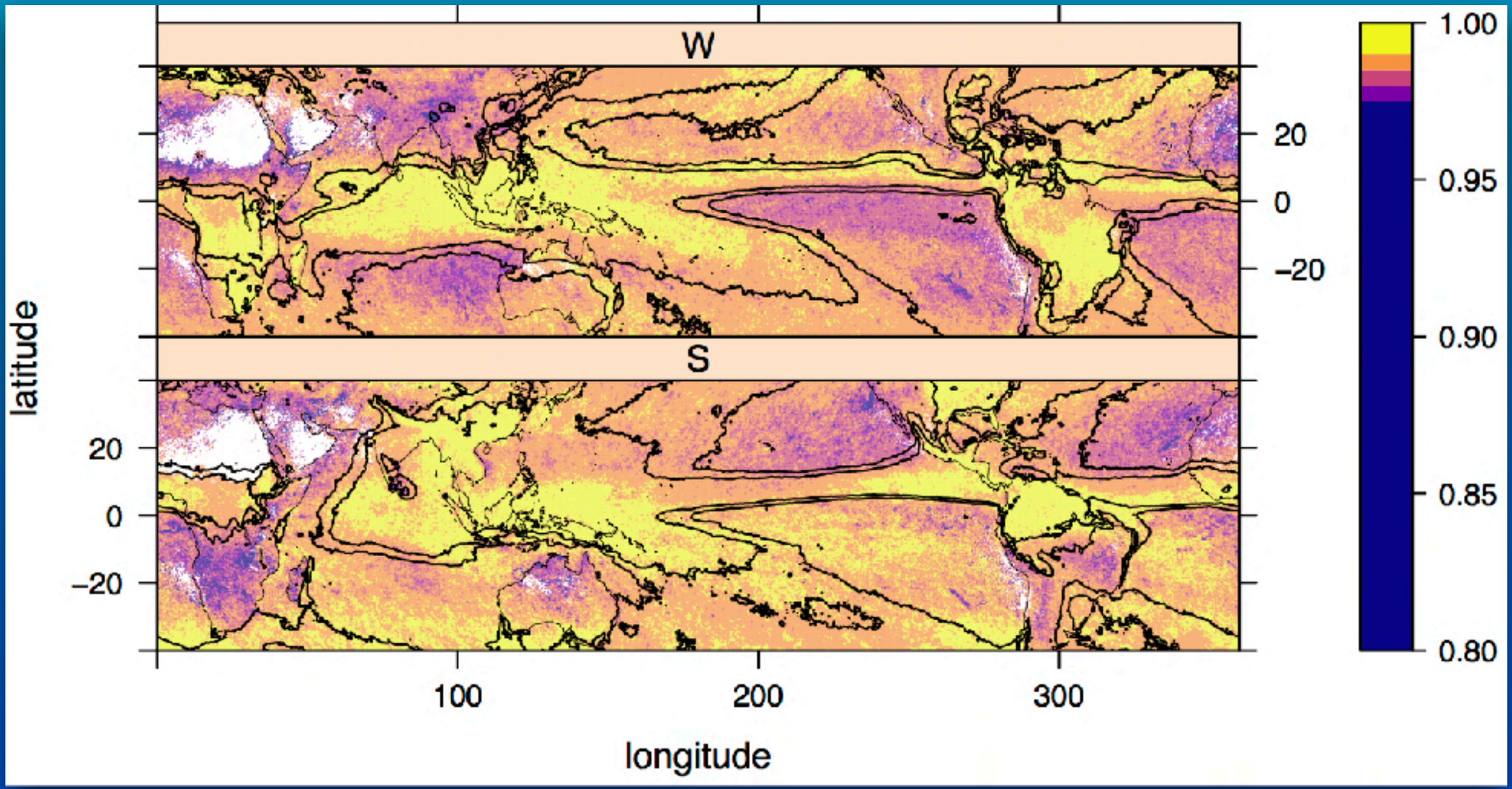
Divide by site-year-season, apply DFA, recombine statistically by site-season.



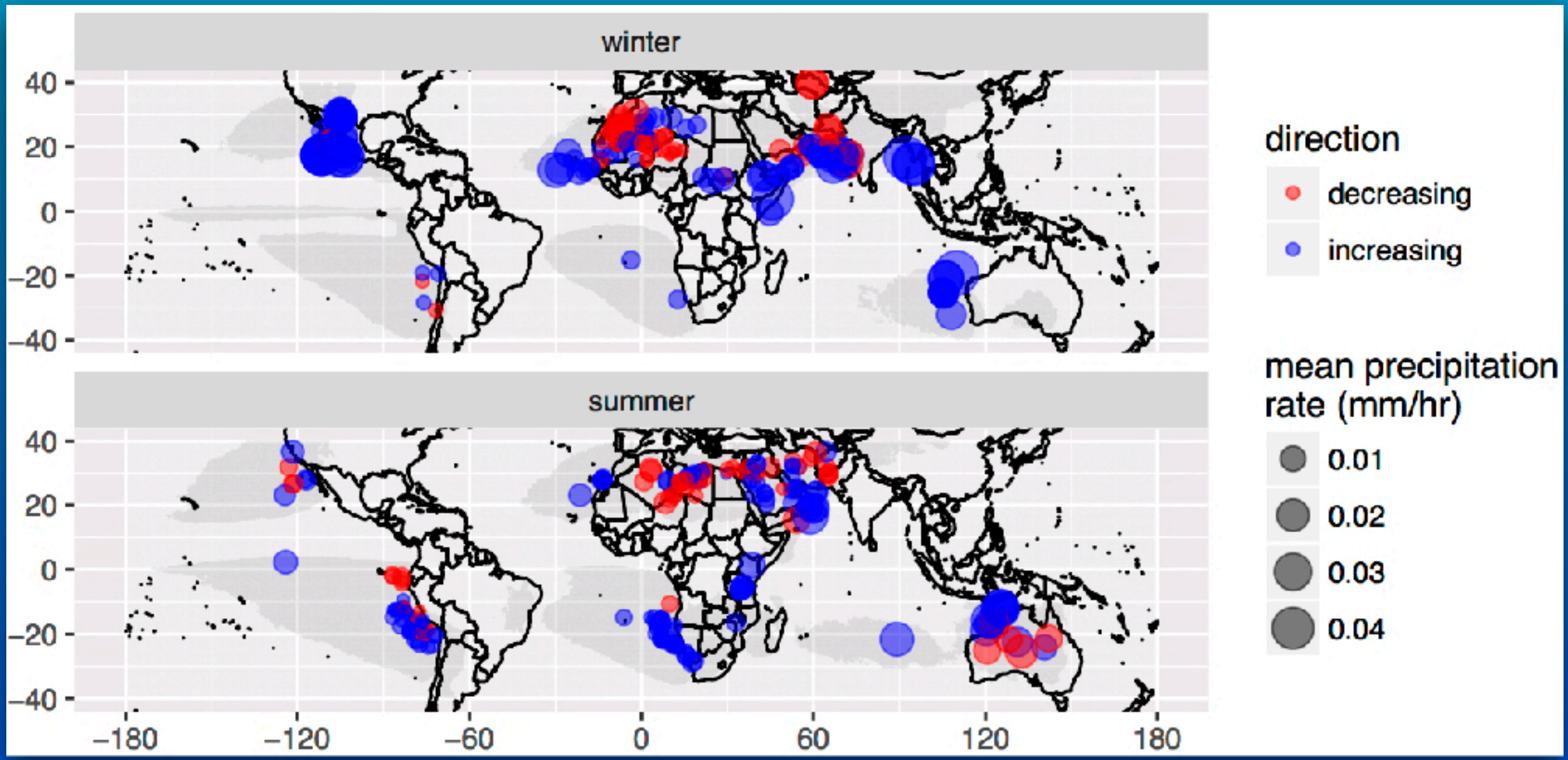
Regional features of H (for up to a month) emerge after statistical recombination

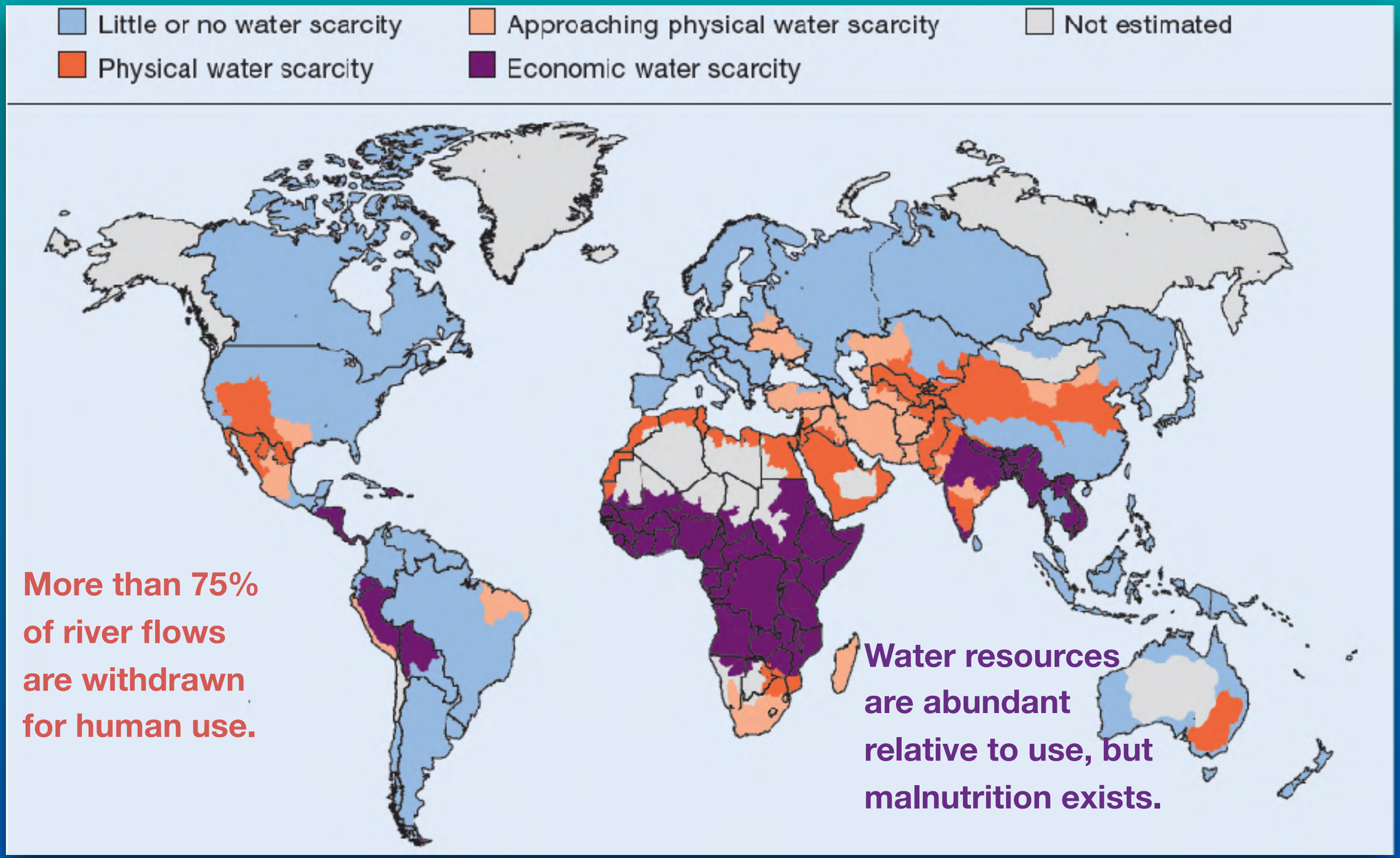


Regional features of fit quality emerge after statistical recombination



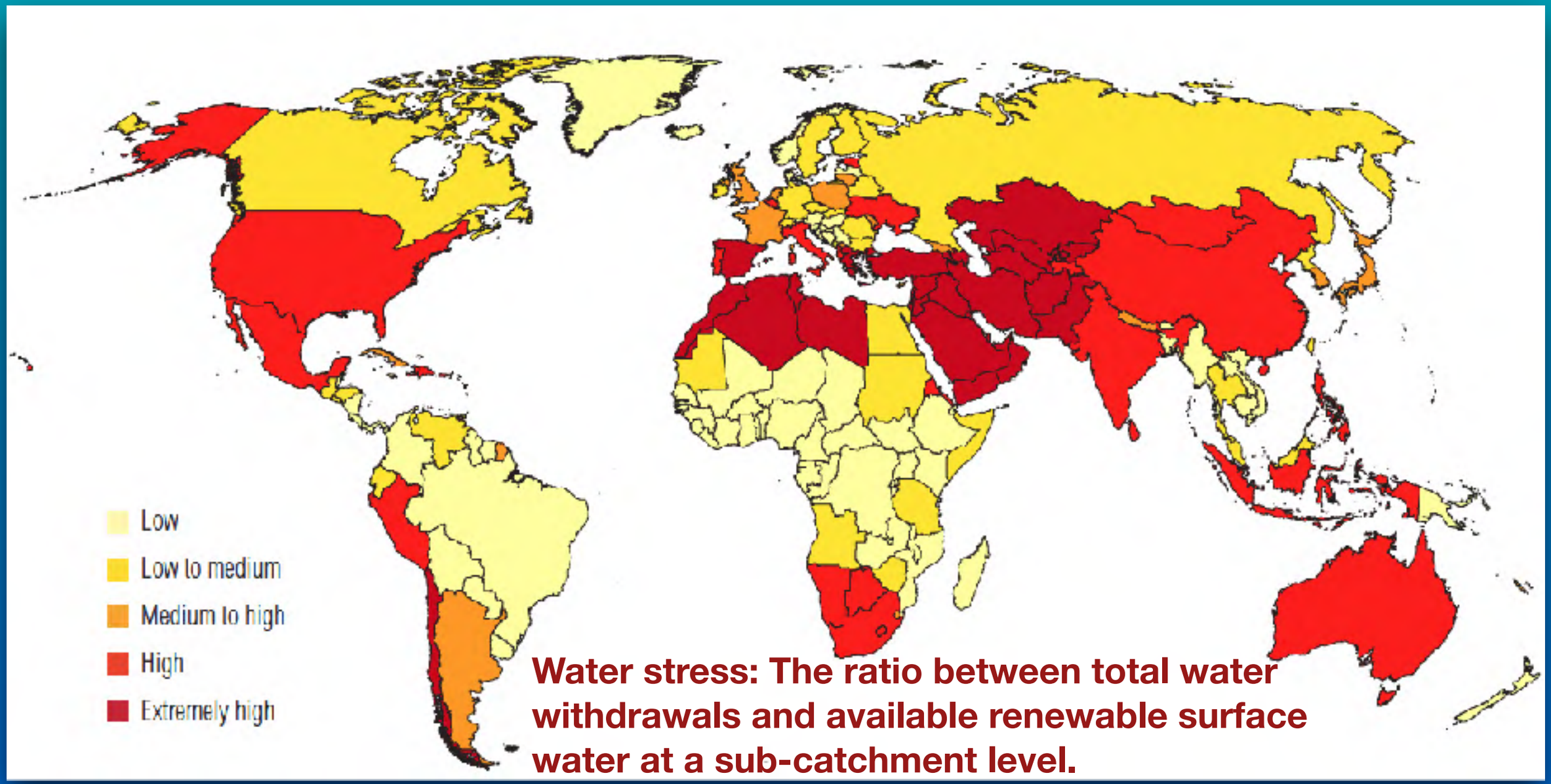
Locations with time change of H over 1998-2015 greater than 0.02 or less than -0.02 (per year)





Areas of **Physical** and **Economic** Water Scarcity

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



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

Summary and Conclusions

- A Data Science project starts with data but ends with concepts, decision-support information, and often useful models
- Divide and recombine enables deep analysis of large complex data
- Hadoop scales D&R to arbitrarily large datasets
- TRMM study:
  - Majority of deep convective precipitating areas in the tropics likely have persistent (temporally highly clustered) precipitation up to a month's time scale
  - Persistent precipitation conditioned by seasons and geography are observed
  - Changes of persistence in precipitation are observed at water scarce and stressed regions from 1998 to 2015—important information for water resource management

Divide & Recombine (D&R)
with the DeltaRho D&R Software
D&R + $\delta\rho$

Meeting the statistical and computational challenges of

- Deep analysis of big data
- High computational complexity of analytic methods

http://deltarho.org

The D&R Framework: Choices of the Analyst

Statistical Division Method: [D] Operations

- divides data into subsets
- division persists, used for many analytic methods

Analytic Method: [A] Operations

- applied to each subset, resulting in subset outputs
- no communication among the subset computations
- embarrassingly parallel: simplest parallel computation

Statistical Recombination Method for Each Analytic Method: [R] Operations

- statistical recombination method applied to outputs
- this is the D&R result for the analytic method
- often has embarrassingly parallel component

DeltaRho software implements D&R

DeltaRho Front End: datadr R package

The analyst programs in R and uses the datadr R package

datadr is a domain specific language for D&R

First written by Ryan Hafen at PNNL (former grad student in Purdue Statistics)

1st implementation Jan 2013

Analyst R and datadr code specifies divisions, analytic methods, and recombinations

DeltaRho Back End

Principal back end so far has been Hadoop

Runs on a cluster

MapReduce: distributed parallel compute engine

HDFS: Hadoop Distributed File System

DeltaRho RHIPE

R and Hadoop Integrated Programming Environment

Provides communication between datadr and Hadoop

Also provides programming of D&R but at a lower level than datadr

First written by Saptarshi Guha while a grad student in Purdue Statistics

1st implementation Jan 2009

DeltaRho Back End: What does Hadoop Do?

Runs the analyst's R code for divisions [D], analytic methods [A], and recombinations [R] in parallel on a cluster of nodes (servers)

Writes subsets and outputs to the HDFS, spreading them across all cluster nodes

- R data structures
- specified by the analyst R + datadr code

DeltaRho Back End: What does Hadoop Do?

Schedules computations: assigns a cluster core to a subset or output on the HDFS to run R code

One of the clusters used by our D&R research team has 10 nodes with 22 cores each, 220 cores in all

There can be 1,000s to 1,000,000s of subsets and outputs for a big data analysis

Maximum of 220 subsets or outputs can run at the same time

R code for subsets and outputs is run sequentially

When the R code computation for a subset finishes, Hadoop assigns another subset to the core

Subject-Matter Division

It is natural to divide data based on the subject matter

Divide by conditioning on the values of variables important to the analysis

Just as valid for small datasets

- widely practiced in the past
- a statistical best practice

D&R with DeltaRho takes advantage of this best practice for computational gain

There are other of division categories, but subject-matter division is the the most used in practice

Satellite Data: Tropical Rainfall Measuring Mission

Part of collaborate work of Purdue Statistics and Earth & Atmospheric Sciences Departments

Version 7, 3B42: 50,632 3-hr rainfall measurements at each of 576,000 locations

Division (1) By Time Across Locations
50,632 subsets, 576,000 measurements per subset

Division (2) By Location Across Time
576,000 subsets, 50,632 measurements per subset

Division (3) Division (2) but in addition each location broken into two subsets, May-October and November-April

This and other examples will be discussed by Professor Tung

Get DeltaRho Software and Documentation

Github is the development site: github.com/delta-rho

Get code and documentation: deltarho.org

Open source with both a GPL and Apache license

Available for download from R CRAN for installation on a cluster

Appliance to spin up a cluster on the Amazon AWS service

Much documentation for datadr and RHIPE

Another distribution site? Started incubation process to become a project of the Apache Software Foundation.

We Invite You to Join Us

You can be a user, a software contributor, or collaborator with us at Purdue

User

- download code, install, and ask questions on Github

Contributor

- visit our Github Organization page
- check out our contributing guide
- feel free to fork any of our component repositories
- datadr , Trelliscope , RHIPE
- introduce yourself on our gitter dev chat room

Collaborator

- Contact Bill Cleveland
- wsc@purdue.edu

D&R with DeltaRho
Analysis and Statistical Modeling of Rainfall Rate
Version 7, 3B42 Tropical Rainfall Measuring Mission

Qi Liu
Statistics

Vinayak Rao
Statistics

Wen-wen Tung
Earth, Atmospheric, & Planetary Sciences

Bill Cleveland
Statistics & CS

Purdue University

TRMM Data

3-hourly rain rate (mm/hr)

1998-01-01 to 2015-04-30

Latitude: 50°N-S

Longitude: 0°-360°

Spatial resolution 0.25 deg x 0.25 deg

576,000 locations

50,632 observations/location

2 Divisions: by time and by location

Build a Bayesian Statistical Model for the Rain Rates

We carry out extensive diagnostic methods to verify that the model fits the patterns in the data

Deep analysis: Analyze data at their finest granularity, in detail, and not just summary statistics

We do not just drop a model on the data and hope for the best.

Build a Bayesian Statistical Model for the Rain Rates

Many visualization methods are used

Apply a method to each subset in sample of subsets

The number of subsets is typically too large to look at plots of all of them

Sampling plans are rigorous because we can readily compute variables of a plan across all subsets

D&R enables all of this

The Statistical Model Building Framework

(1) Exploratory visualization

(2) Candidate model based on exploration and subject matter knowledge

(3) Fit the model to the data

(4) Model diagnostics

(5) If model fails diagnostics, typically the visualization tools provide insight on how to alter the model

(6) Return to (2) with the altered model becoming the candidate

(7) Iterate (2)-(6) until model passes

Begin with Study of Missingness

Results

Reduce data

From 50°N-S to 40°N-S

Start 1998-07-01 instead of 1998-01-01

Probability of Rain: Forecast

Logistic regression

Response is rain or not for each location L at time T

Spatial explanatory variables are rain or not for time T-1 at the 8 neighbors, N, of L and L itself

```
L = location and N = neighbor

      N     N     N

      N     L     N

      N     N     N
```

Other explanatory variables: hour of day, month of year

The fitted logistic regression provides a probability of rain for each L

We can get a confidence interval for the probability, say 95%, from the Bayesian posterior distribution.

Probability of Rain: For Understanding

L is again the response at time T, rain or not

But now the current values, rain or not, of the 8 neighbors are also explanatory variables, as well as the 9 values for time T-1

Other explanatory variables: hour of day, month of year

We can get a confidence interval for the probability, say 95%, from the Bayesian posterior distribution

Down sampling ???