Divide and Recombine (D&R) Data Science Projects for Deep Analysis of Big Data and High Computational Complexity

By:

W. Tung, M. C. Bowers, A. Barthur, Y. Song, W. S. Cleveland
Purdue University

J. Gerth
Stanford University

Technical Report #18-01

Department of Statistics
Purdue University

March 15, 2018

Divide and Recombine (D&R) Data Science Projects for Deep Analysis of Big Data and High Computational Complexity

W. Tung and M. C. Bowers
Purdue Department of Earth,
Atmospheric, and Planetary Sciences
A. Barthur
Purdue Center for Education and Research
in Information Assurance and Security
J. Gerth
Stanford Departments of Computer
Science and Electrical Engineering
Y. Song and W. S. Cleveland
Purdue Department of Statistics

Abstract

The focus of data science is data analysis. This article begins with a categorization of the data science technical areas that play a direct role in data analysis. Next, big data are addressed, which create computational challenges due to the data size, as does the computational complexity of many analytic methods. Divide and Recombine (D&R) is a statistical approach whose goal is to meet the challenges. In D&R, the data are divided into subsets, an analytic method is applied independently to each subset, and the outputs are recombined. This enables a large component of embarrassingly-parallel computation, the fastest parallel computation. **DeltaRho** open-source software implements D&R. At the front end, the analyst programs in **R**. The back end is the **Hadoop** distributed file system and parallel compute engine. The goals of D&R are the following: Access to the 1000s of methods of machine learning, statistics, and data visualization; Deep analysis of the data, which means analysis of the detailed data at their finest granularity; Easy programming of analyses; High computational performance. To succeed, D&R requires research in all of the technical areas of data science. Network cybersecurity and climate science are two subject-matter areas with big, complex data benefiting from D&R. We illustrate this by discussing two datasets, one from each area. The first is the measurements of 13 variables for each of 10,615,054,608 queries to the Spamhaus IP address blacklisting service. The second has 50,632 3-hourly satellite rainfall estimates at 576,000 locations.

1 What is Data Science?

Data science is, at its foundation, centered on the analysis of data. The technical areas of data science are those that need direct study to make data analysis as effective as possible (Cleveland, 2001; Cleveland and Hafen, 2014). The areas are:

- statistical theory,
- statistical models,
- statistical and machine-learning methods,
- visualization methods,
- algorithms for statistical, machine-learning, and visualization methods,
- computational environments for data analysis: hardware, software, and database management, as well as
- live analyses of data where results are judged by the subject-matter findings, not the methodology and systems that are used.

Of course, these areas can be divided into sub-areas, which in turn can have sub-sub-areas, and so forth. Also, research in an area can depend heavily on research in others.

1.1 The Role of Live Data Analyses

Researchers in an organization who work on one or more of the of areas (1) to (6) need to have exposure to live data analyses that are an integral part of the organization's culture. "Live" means the success of the analysis is judged by the subject matter results. Live analyses serve as an effective heat engine for research ideas that solve real problems, and serve as a testbed for the those ideas (Cleveland, 2001, 2005; Cleveland and Hafen, 2014). Analyses need to be live because it is important to judge research based on how well it increases subject-matter knowledge and the effort required to get the increases. There are notions of statistical accuracy that can be important, too, but those become subject to the same criterion of increasing subject-matter knowledge.

So, live analysis of data enters into a symbiosis with the technical areas of data science. This is illustrated by the monumental contribution of John W. Tukey to numerical spectrum analysis (Brillinger, 2002). His close ties with physical scientists and their data stimulated him to lay a foundation for power spectrum analysis. He provided many tools such as the Fast Fourier Transform and the simple idea of tapering to decrease leakage. This led to substantial increases in the resolution of spectrum estimates. The data-driven work, in turn, led to deeper understanding of physical science phenomena and discoveries such as the normal modes of the Earth, and the modes of atmospheric waves.

1.2 Computational Environments for Data Analysis

One way to think through environments for data analysis is to divide software into the front end and the back end, possibly with software that connects the two. The front end is software that a data analyst uses to specify commands for analysis. The back end is software that executes analytic routines.

What we want from the front end is software that is very time-efficient for the data analyst. A good way to judge this is to compare (1) the fraction of time the analyst spends programming with the data to carry out the analysis, and (2) the fraction of time spent thinking about the subject matter and about the choice of analysis methods to use to maximize what is learned. Clearly, we want (2) to be as large as possible and (1) as small as possible. We can describe this as easy to program with the data, but this is not really enough. We want the front end to be a language for data analysis that is also powerful, allowing analysis of the data at their finest granularity. We call this "deep analysis". Analyzing summary statistics is important to analysis, but doing only that and not studying the data in detail runs the risk of missing important information in the data.

One example of an excellent front end is the **R** language for data analysis. It is the open-source implementation of the **S** language, which won the ACM System Award in 1998. It was developed at Bell Labs. The citation was: "John M. Chambers. For The **S** system, which has forever altered how people analyze, visualize, and manipulate data." Other past award winners included Unix, Visicalc, and the World Wide Web. So, **S** has quite good company.

S and therefore **R** provide very efficient and powerful programming. To do this, computational performance was slower than some other systems. The reason is that whenever a choice had to be made between efficiency of the user or efficiency of the computation, the user always won. However, today, this is a less critical matter because the only way to face big data with high computational complexity is through parallel computing. This has been done with **R**, and will be described below. Another important aspect of **R** is that it has by far the most implementations of analytic methods of any system for data analysis. Currently, the **R** CRAN repository has 12,179 packages along with many analytic methods in the core distribution.

2 Divide & Recombine (D&R) with DeltaRho

2.1 D&R

Computational performance is challenging today. Datasets can be big, computational complexity of analytic methods can be high, and computer hardware power can be limited. Small datasets can be challenging, too, when the computations have high complexity. Divide & Recombine (D&R) is a statistical approach to meet the challenges (Guha et al., 2009, 2012; Cleveland and Hafen, 2014).

In D&R, the analyst divides the data into subsets by a D&R division method. Each analytic method is applied to each subset, independently, without communication. The outputs of each analytic method are recombined by a D&R recombination method. Sometimes, the goal is one result for all of the data, such as a logistic regression; D&R theory and methods seek division and recombination methods to optimize the statistical accuracy. Much more common in practice is a division based on the subject matter. The data are divided by conditioning on variables important to the analysis.

2.2 DeltaRho Software for D&R

D&R computation is mostly embarrassingly parallel, the simplest parallel computation. The **DeltaRho** software is an open-source implementation of D&R (`http://www.deltarho.org`, Hafen et al. 2016.) The front end is the **R** package **datadr**, which is a domain specific language for D&R. The data analyst programs in **R**, and uses **datadr** to specify divisions, applications of analytic methods, and recombinations. The **datadr** package makes programming D&R simple. At the back end, running on a cluster, is a distributed database and parallel compute engine such as **Hadoop**, which spreads subsets and outputs across the cluster, and executes the analyst's **R** and **datadr** code in parallel. The **R** package **RHIPE**, the **R** and **Hadoop** Integrated Programming Environment, provides communication between **datadr** and **Hadoop**.

2.3 Division

Subject Matter Division. It is natural to divide data based on the subject matter. The data are divided by conditioning on variables important to the analysis. For example, the Tropical Rainfall Measuring Mission (TRMM, Huffman et al. 2007) multi-satellite dataset analyzed in Sec. 4 initially has 50,632 3-hourly rainfall records at each of the 462,240 locations. There are missing values. One typical division is by location across time, so there are 462,240 subsets with 50,632 records per subset. Another is by time across locations, so there are 50,632 subsets with 462,240 locations per subset.

Subject matter division is just as valid for small datasets. It has been widely practiced in the past and is a statistical best practice. For D&R we use such division both for a best practice and for computational gain. Subject-matter division is the most used in practice.

Sampling Division. Sometimes the goal is one result for all of the data, for example, a logistic regression. In this case, each subset in a division is seen as a sample of the data. Subsets are replicate samples. For example, we can carry out random replicate division: choose subsets randomly. Another division method is to choose subsets to make each as representative of the data as possible. The D&R result is not the same as that we could have gotten had we been able to apply the logistic regression to all of the data. However, D&R theory and methods seek division and recombination methods to optimize the statistical accuracy and make its accuracy as close as possible to that of the all-data estimate.

How fast? Logistic Regression for a 1 TB Dataset. We carried out logistic regression with 2^{30} observations of the response and 127 explanatory variables. There is 1 TB of data. There were 2^{20} subsets, each with 2^{10} observations per subset. The cluster had 11 nodes with 264 cores, 528 GB total RAM, and 88 TB total disk.

We measured the elapsed time to read subsets into memory and form the subsets. We also measured the elapsed time to carry out the computation of the logistic regressions of subsets and the recombination to get a single result. The two time measurements were 12.1 min and 6.0 min, respectively, a total of 18.1 min. This is not an undue amount of time in practice.

2.4 D&R: What do we get?

Deep Analysis. We get deep analysis, as described above, even when the data are big and computational complexity is high. This includes visualization of the detailed data, critical to statistical model building and validation, as well as to determining if a machine learning method is appropriate for the visualized patterns in the data. We do visualization by applying a method to subsets that have the detailed data. While it is feasible typically to apply a visualization method to all subsets, it is often not practical to look at them all because there can be far too many subsets, which in applications can be tens of thousands to the millions. So, we sample. Sampling plans that preserve the information in the data can be readily devised because we can compute sampling variables across all subsets to use as the basis of a rigorous sampling plan.

High Computational Performance. We get high computational performance. **DeltaRho** can increase dramatically the data size and analytic computational complexity that are feasible in practice, whether hardware power of an available cluster is small, medium, or large. The data can have a memory size that is larger than the physical cluster memory. For us, this occurs routinely. Approaches to big data analysis that rely on data to be stored in memory to get good computational performance, for example, because there are iterations, place a heavy limitation on datasets that can be analyzed.

Access to Methods. We get access to the thousands of methods of statistics, machine learning, and data visualization. Of course, **R** provides this.

High Efficiency Programming by the Analyst. We get very high efficiency in using **R** and **datadr** to program with the data, along with great power and flexibility that allow deep analysis and tailoring analyses to the data. Most importantly, **DeltaRho** protects the analyst from the detail of distributed parallel computation and subset database management. Furthermore, **datadr** is abstracted from the back-end choices, so that its code is the same whatever the back end. For example, you can use **datadr** on a single multicore machine. Of course, back ends other than **Hadoop** require other software that connects **datadr** and a back end like **RHIPE** does for **Hadoop**.

The Price is Excellent. It's free. **GitHub** is the development site: `http://github.com/delta-rho`. The software is open source with both GPL and Apache licenses. It is available for download from **R** CRAN for installation on a cluster. We have an appliance to spin up a cluster on the Amazon AWS service. We have much documentation for **datadr** and **RHIPE**. To get the software and documentation, please see `http://www.deltarho.org`.

2.5 Computational Performance Measurement and Analysis (CPM&A)

In the world of big data and high computational complexity, CPM&A is critical but is often lacking in rigor, and not sufficiently informative. Research for CPM&A can greatly enhance the performance of data analysis. It can also compare rigorously different computational methods and

systems. Conventional performance tests, however, use only a few low-level computations such as sort. Such "benchmarks" are not informative for data analyses. Instead, CPM&A should be the study of elapsed times of analytic methods, which are directly what an analyst experiences. Also, benchmark testing today often fails to control for factors that are important for comparing aspects of two systems. For example, configuration parameters for **Hadoop** and its competitor **Spark** can have a big impact on performance. Pilot experiments have shown that there are strong interactions among factors (e.g., Shi et al., 2015). Little attention today is given to factor interactions. These problems must be addressed by multifactor experiments that include interactions when needed.

Experimentation is challenging. Knowledge of big-data systems is needed to determine salient factors. Interactions are not readily quantifiable from the knowledge. So, empirical model-building is necessary. Running a design on a high-performance computing cluster can be complex. Some factors have levels that can be changed only by system administrators. In comparing **Hadoop** and **Spark** it is important to keep factors fixed for each system, and both must be installed on a cluster. Attention must be paid to limiting all other processes other than the operating system. Care must be taken to ensure replicate runs are independent and not affected by aspects of the systems that can create trends in duplicate runs. Many topics need to be evaluated by running on more than one cluster.

3 Example 1: Spamhaus Blacklist IP Address Data

3.1 Introduction

Each device connected to the Internet has an IP address. An Internet connection consists of one IP address, the source, seeking to connect to another, the destination. The purpose is to achieve transmission of information that is broken up into packets whose sizes are up to 1500 bytes. On the way are a series of devices with IP addresses that forward the packets from one device to the next along a path from source to destination. The Spamhaus blacklisting service (`http://www.spamhaus.org`) classifies IP addresses of devices as blacklisted based on reports of device behaviors such as being a major producer or conduit of spam. Many network operations use this service. For example, email servers check IP addresses of the source and previous forwarding devices of an email message by sending "queries" about all the IP addresses to a blacklisting service. The "response" says whether each queried IP address is blacklisted or not, and if so, the categories of network misuse that occurred, for example, spam. If any device is blacklisted, the email server can decide to block the message and not forward. One aspect of blacklisting is that it is dynamic; a queried IP address can vary between blacklisted and not-blacklisted.

We collected the query/response data for 10,615,054,608 queries from the Stanford mirror of the Spamhaus service. The collection period was 18 weeks. Each query results in values of 13 variables. The total memory size of the data, which consists of R objects, is 3.71 TB. The cluster on which we analyze the data has 1.28 TB of physical memory. This provides an example of our statement in Sec. 2 that needing data to reside in memory to get good computational performance is very limiting.

We have addressed a number of topics in analyzing the data. One topic is to study through time, all

queries of each unique queried IP address that has at least one blacklisting result in its queries; there are 59,435,635 such blacklisted addresses. Altogether, there are 207,081,108 queried IP addresses, so 29% are blacklisted at least once. The largest number of queries for the addresses is about 1.5 million. The distribution is very skewed. About 80% of the addresses have less than 16 queries.

The division of the data has each of the 59,435,635 blacklisted IPs as a subset, so it is a subject-matter division. Each subset consists of values of 9 variables at the query points in time for each blacklisted IP through time. One variable is the timestamp of each query. The other 8 variables provide information about the query. One is the IP address of the querying host. Five violation variables specify whether each was violated or not. If there is at least one violation, then the query result is a blacklisting at the query timestamp. If none are violated, the result is not-blacklisted at the timestamp.

The result for each blacklisted IP address is a marked point process. The timestamps are the points and the other 8 variables are the marks. So each subset is a time profile for the blacklisted IP address. The profiles are marked point processes whose number of points, a very skewed distribution, ranges from 16 to 2,097,152. An **R** dataframe object for a blacklisted IP address has 9 columns and each row is a query. The timestamps have runs of blacklistings (on's) and runs of not-blacklistings (off's), creating an on-off process which is of great interest.

3.2 Blacklisting and Not-Blacklisting Rates

Spamhaus has procedures for removing a blacklisting. But, they are complex and require verification by the Spamhaus staff. If a highly reputable Internet service provider reports that one of their blacklisted IP addresses assigned to a customer should be unlisted, then Spamhaus does so. Or, Spamhaus does its own investigations. An individual or organization can use a tool provided by Spamhaus to remove a blacklisting. This is a slow manual process, however. One would expect that blacklistings would persist for some time. One very interesting topic then is the blacklisting and not-blacklisting marked point process for a queried IP. Studying this comprehensively is very complex, and beyond the scope of this article. Here, we will show a few suggestive properties that need much further investigation.

Figure 1 is a plot for one blacklisted IP address with $512,783$ queries over a period of 96 days. There are $271,986$ blacklistings and $240,797$ not blacklistings, close to 1/2 each. The queries were broken up into blocks of 2^{12} consecutive queries through time. The last time for each block is the first time of the next block to include all intervals between queries. There are 125 intervals. For each block, the duration from the first query to the last in min is divided into 2^{12} to get a number of queries per min. Also, for each block, the number of blacklistings and not-blacklistings were divided by the duration to get rates in number of listings per min. In Fig. 1, the rates of blacklistings and not-blacklistings are plotted against query rates on a log base 2 scale for the 125 blocks. A line was fitted to the data for the blacklisted and for the not-blacklisted using robust regression. The intercept and slope are 0.3528 and 0.8093, respectively, for not-blacklisted, and 0.2950 and 0.8347 for blacklisted.

Figure 1 suggests there may be a rapid back and forth of the on-off process, the blacklisting runs the not-blacklisting runs, rather than long periods of each. Across all rates, the fractions of black-

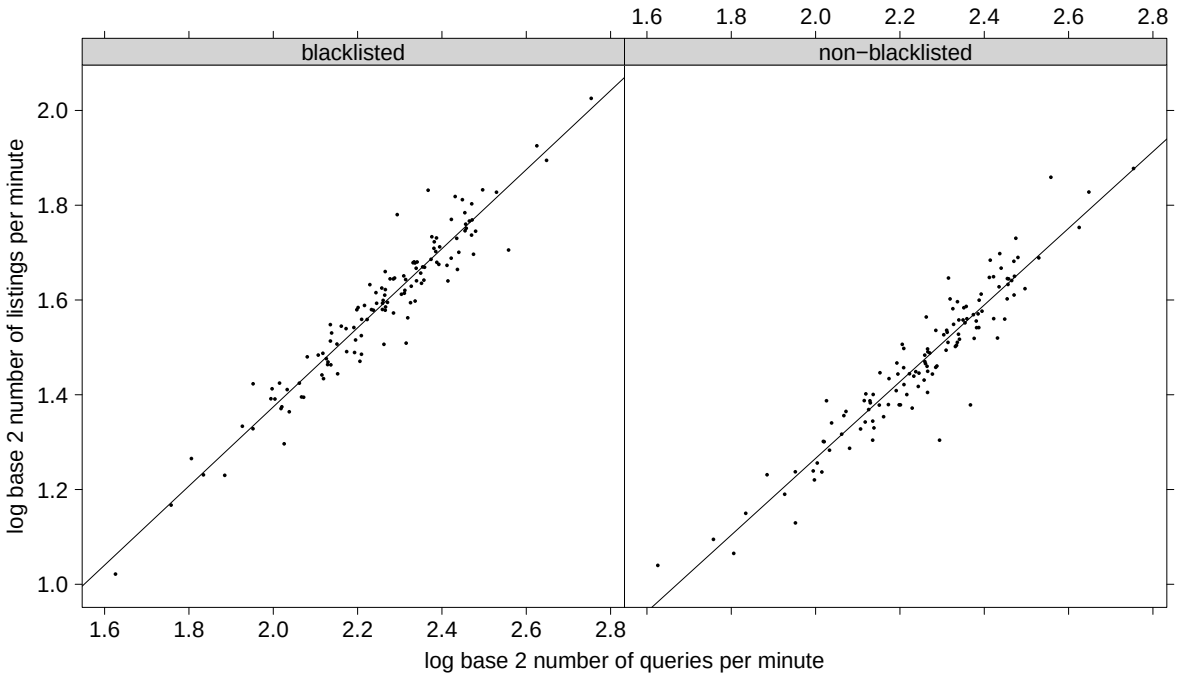

Figure 1: *Listing rates, blacklisted and non-blacklisted, are plotted against query rates. A line is fitted to the data in both cases using robust linear regression.*

listings and not-blacklistings are close to the overall fractions, about 0.5 each. The blacklisting and not-blacklisting rates are not far from being proportional to the query rates. The duration of the largest query rate block is about 9.7 hr. It has 2340 blacklistings and 2153 not-blacklistings. Understanding the on-off processes generally will have to come from further deep analysis of the blacklisted IP marked point processes.

4 Example 2: Large Complex Weather and Climate Data

4.1 Background: Science Meeting Societal Impacts

Climate change, population growth, and extreme weather are increasingly challenging the limited natural resources of Earth. For society to understand and counter this challenge, it is critical to integrate multi-disciplinary knowledge at the nexus of intricately coupled systems: weather, climate, environments, resources, energy, and societies. This is exemplified by the integrated framework of the United Nations Sustainable Development Goals (`http://Sustainabledevelopment.un.org`). van Vliet et al. (2016), for example, studied climate change and the resulting changes in water resources available to power generation. Hydropower and thermoelectric power-generation technologies comprise 98% of the world's electricity generation. Both strongly depend on water availability; thermoelectric technology further depends on water temperature for cooling. van Vliet et al. (2016) suggested that as the flow of many rivers is projected to decrease while water tem-

perature rises, in future decades more power plants will be forced to shut down or lose efficiency, presenting uncertainty in other energy sectors as well as regional economy and political stability. Developing nations rely heavily on these power-generation methods and, therefore, will likely face more of these setbacks than developed ones.

The World Economic Forum has been conducting annual Global Risk Perception Surveys among international multi-stakeholder communities since 2007 (e.g., World Economic Forum, 2017). In recent years, extreme weather events have been among the top five global risks perceived as mostly likely to occur. In the Forum's 2017 Global Risk Report, it has risen to be the most likely to occur and the second highest impact (after weapons of mass destruction). Failure of climate change mitigation and adaptation is also among the top five highest-impact risks by perception. The risks posed by extreme weather events, failure of climate change mitigation and adaptation, food crises, water crises, and large-scale involuntary migration are perceived to be strongly linked. Yet, each of these risks is influenced by trends such as changing climate, degrading environment, or rising urbanization. These perception surveys have motivated numerous data-driven research.

Earth's complex natural and human-built systems are challenging but fertile application areas for data science. Generating insights and knowledge with data-driven research in complex global issues requires geoscientists having access to the big datasets that must be collected for scientific study of the systems, and access to computational environments for data analysis that enable deep analysis of the data, high computational performance, and highly efficient programming with the data. This understanding has been manifested in prior data-driven research, in part summarized in (Sellars et al., 2013) such as in automated evaluation of satellite-based estimates of tropical rainfall using artificial neural networks (Sorooshian et al., 2000), object-based approaches for verification of precipitation forecasts (Davis et al., 2006a,b) or to study the impact of climate variability on precipitation (Sellars et al., 2015), interactive 3D visualization of geospatial data (Mitasova et al., 2012), and machine-learning techniques to diagnose aviation turbulence (Williams, 2014) or to extract basic modes of tropical atmospheric convection (Tung et al., 2014).

NASA's Earth Observing System Data and Information System represents the US federal government's system-wide investment to innovate open data access. As a result, scientists from multiple disciplines are able to study Earth systems collaboratively in a data-driven environment. Scalable data analysis techniques have become ever more critical in anticipation of the deluge of current and future satellite data. Following, we present the analysis of $\sim$17 years of global tropical precipitation on fixed $0.25^\circ \times 0.25^\circ$ grids and 3-hourly time intervals from the Tropical Rainfall Measuring Mission. In the analysis, we calculated the long-range dependence (long-memory) property of precipitation at each grid point (location) of the data with **DeltaRho**.

4.2 TRMM Dataset

Precipitation, including the process of precipitation, has both operational and fundamental scientific importance. It is the part of the local and global water cycle that concentrates the heat used to evaporate the water that is transported in the atmosphere, and it delivers water to the ocean or land surfaces.

The Tropical Rainfall Measuring Mission was a joint mission of NASA and the Japanese Aerospace

Exploration Agency (Simpson et al., 1996). The TRMM Version 7 3B42 Multi-satellite Precipitation Analysis (TMPA) data product combines calibrated precipitation estimates from multiple satellites and surface rain gauges where feasible (Huffman et al., 2007). The dataset has global spatial coverage from 50° S–50° N with $0.25^\circ \times 0.25^\circ$ horizontal resolution, spanning the period from 1998 through early 2015 with 3-hourly resolution. At fine time scales, the TMPA successfully reproduces the surface observation-based distribution of precipitation as well as large daily events. However, in common with other fine-scale rainfall estimators, it has lower skill in correctly specifying low and moderate rainfall amounts on short time scales (Huffman et al., 2007). We removed data before October 1998 and pole-ward of 40° latitudes because of a nontrivial fraction of missing data. The small fraction of remaining missing data points were imputed by linear interpolation. The total data size was about 1/4 TB, not very big. Yet, in this project we were challenged by the computational complexity of the analytic methods.

4.3 Detrended Fluctuation Analysis

The persistent temporal scaling properties of a time series can be conveniently characterized by the Hurst exponent or Hurst parameter H (Hurst, 1951). The Hurst parameter quantifies the persistence of correlations such that when $0 < H < 1/2$, the signal has anti-persistent correlations; when $H = 1/2$, the signal is memoryless or has short-range correlations; when $1/2 < H < 1$, the signal has persistent long-range correlations; and when $H > 1$ the signal may be non-stationary or have non-trivial trends. While the Hurst parameter can be estimated by a variety of methods, few are capable of accurately estimating H when $H > 1$ (Gao et al., 2006, 2007). One of the few methods capable of accurate estimation in the presence of trends or non-stationarity is detrended fluctuation analysis (DFA) (Peng et al., 1994). Because of the potential for non-stationarity associated with tropical expansion, we use DFA to estimate the strength of temporal scaling in the tropical rainfall data.

DFA works as follows: given a noise (or increment) time series, $x_1, x_2, x_3, \cdots$, with mean $\bar{x}$, one first constructs a random walk process,

$$u(i) = \sum_{k=1}^{i} (x_k - \bar{x}), \;\; i = 1, 2, \cdots, N \tag{1}$$

then divides $\{u(i),\ i = 1, 2, \cdots, N\}$ into $\lfloor N/m \rfloor$ non-overlapping segments (where $\lfloor N/m \rfloor$ denotes the largest integer equal to or smaller than N/m), each containing m points. The local trend in each segment is then computed, typically as the ordinate of a best linear or polynomial fit of the random walk in that segment. Finally, one computes the "detrended walk", denoted by $u_m(k), k = 1, 2, ..., m$, as the difference between the original segment (the"walk"), $u(k)$, and the local trend. The fractal behavior is described by the following scaling law

$$F_d(m) = \left\langle \sum_{i=1}^{m} u_m(i)^2 \right\rangle^{1/2} \sim m^H \tag{2}$$

where the angle brackets denote ensemble averages of all the segments.

While it is common to use linear or polynomial regression to estimate the local trend in each segment, this leads to discontinuities or even large abrupt jumps at the segment boundaries. By

using a smooth trend instead of the discontinuous piecewise trend, DFA can better handle non-stationarity or arbitrary trends in the data (Gao et al., 2011). Here, we used the loess smoother (Cleveland and Devlin, 1988) to compute a globally smooth trend $v_w(i)$, where w denotes the span (bandwidth) of the loess smoother. The residual, $u(i) - v_w(i)$, characterizes fluctuations around the global trend, and its variance yields the Hurst parameter H according to

$$F(w) = \left[\frac{1}{N}\sum_{i=1}^{N}(u(i) - v_w(i))^2\right]^{1/2} \sim w^H. \tag{3}$$

To capture the clustering of rainfall events within each 6-month season, we estimated the fluctuation function $F(w)$ on the time scales from 12 hours to 16 days, which correspond with typical "mesoscale" to "synoptic-scale" atmospheric variability. To estimate H, we used the slope coefficient of a linear regression of $\log_2 F(w)$ against $\log_2 w$ (e.g., Bowers et al., 2013).

4.4 Divide and Recombine

To scrutinize the temporal correlation structure in the TRMM precipitation data, we employed D&R with **DeltaRho** using a **Hadoop** cluster. Our primary interest was in the temporal correlation structure of local precipitation. Since precipitation regimes can vary dramatically across locations and seasons, we divided the dataset into subsets, each consisting of a 6-month-long 3-hourly time series of precipitation rates for a particular location. We used the concept of "monsoon years (Yasunari, 1991)" starting with Northern Hemispheric summer as May through October, followed by winter as November through the next April. For each of the 462,240 locations, we had 17 winter subsets (winter 1998 to winter 2014) and 16 summer subsets (summer 1999 to summer 2014), $\sim$ 15 million subsets in total. We applied DFA to the time series data in each subset to estimate the persistence (as discussed in Sec. 4.5) yielding an estimate of H for each subset. We performed several statistical recombinations of the DFA results. After applying DFA, we performed a statistical recombination of the results for each location and season, e.g., taking an average, yielding a map of that statistic for both Northern Hemispheric summer and winter seasons. The **Hadoop** cluster running the **DeltaRho** software stack provided high computational performance for the divisions, application of analytic methods, and statistical recombinations.

4.5 Results and Discussions

Figure 2 shows two example seasonal subset precipitation time series. Both subsets have a seasonal average precipitation rate of about 3 mm day^{-1}; however, the distribution of rainfall throughout the season is markedly different. Whereas the top panel shows a somewhat homogeneous distribution of rainfall throughout the season, the bottom panel shows a strong clustering of rainfall into a relatively short period of time. While the two subsets have the same seasonal average water availability, the subset in the bottom panel may be considered to experience drought, while that in the top panel does not.

The difference in the temporal distribution of rainfall between the two subsets in Fig. 2 is echoed by the estimates of Hurst parameter. The top panel exhibits a homogeneous temporal distribution or

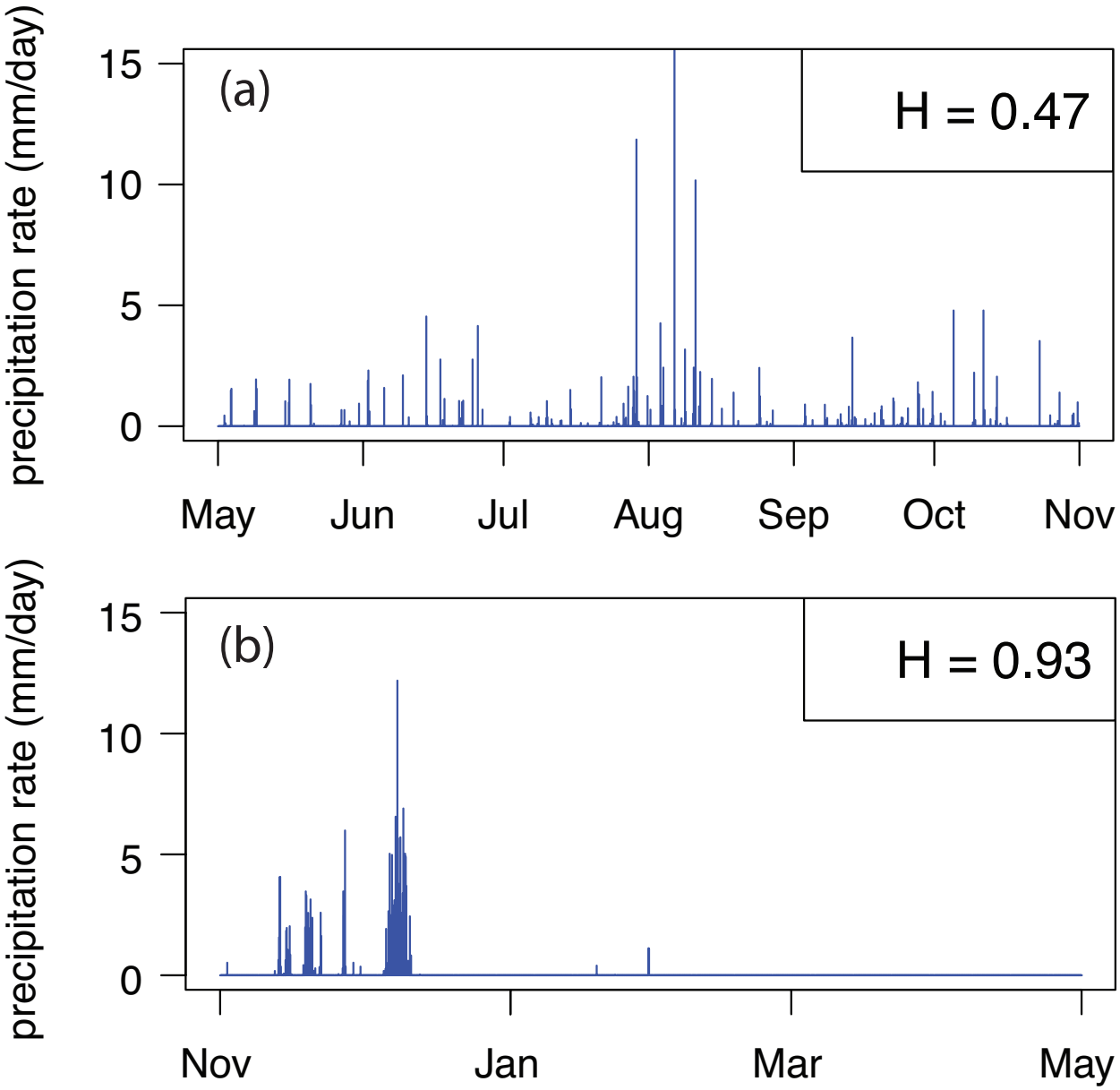

Figure 2: *Seasonal block precipitation time series for (top) Northern Hemispheric summer 2004 at* 32.875°*E,* 7.875°*N and (bottom) winter 2013 at* 85.625°*E,* 14.625°*N.*

weak clustering and has H near 1/2, indicating a memoryless precipitation process from 12 hours to 16 days. On the other hand, the bottom panel exhibits strong clustering and has $H \approx 0.9$, indicating persistence in the precipitation process from 12 hours to 16 days. The persistent clustering in fact exists on multiple scales; on a seasonal scale, the majority of rainfall is confined to the month

from mid-November to mid-December, but on a weekly scale the rainfall within that month is also clustered into distinct events. This fractal behavior is consistent with other findings of self-similarity in rainfall (e.g., Lovejoy and Mandelbrot, 1985). We note that while these two locations have equivalent seasonal average rainfall, the difference in temporal clustering patterns echoed by the Hurst parameter has profoundly different implications for local water resource management.

Having established that the Hurst parameter can distinguish various rainfall clustering patterns we now examine the average spatial distribution of clustering patterns in the global tropical belt. Figure 3 shows the average Hurst parameter for each location and season. Each pixel in the figure represents the average of the 17 winter or 16 summer H values estimated from 12 hours to 16 days at each location. In general, H values range between 1/2 and 1 indicating that tropical precipitation is either memoryless or persistent on the meso- and synoptic- time-scales. There is a pronounced land-sea contrast with a tendency for marine rainfall to be persistent and land rainfall to be memoryless. Persistence tends to be enhanced in regions with highly active atmospheric convection such as the Indian Ocean–Pacific "warmpool", the Inter-Tropical Convergence Zone, the South Pacific Convergence Zone, the tropical Atlantic, and the Caribbean. Persistence also tends to be particularly strong in regions with sharp horizontal gradients in mean precipitation, as indicated by the 2 and 4 mm day^{-1} contours.

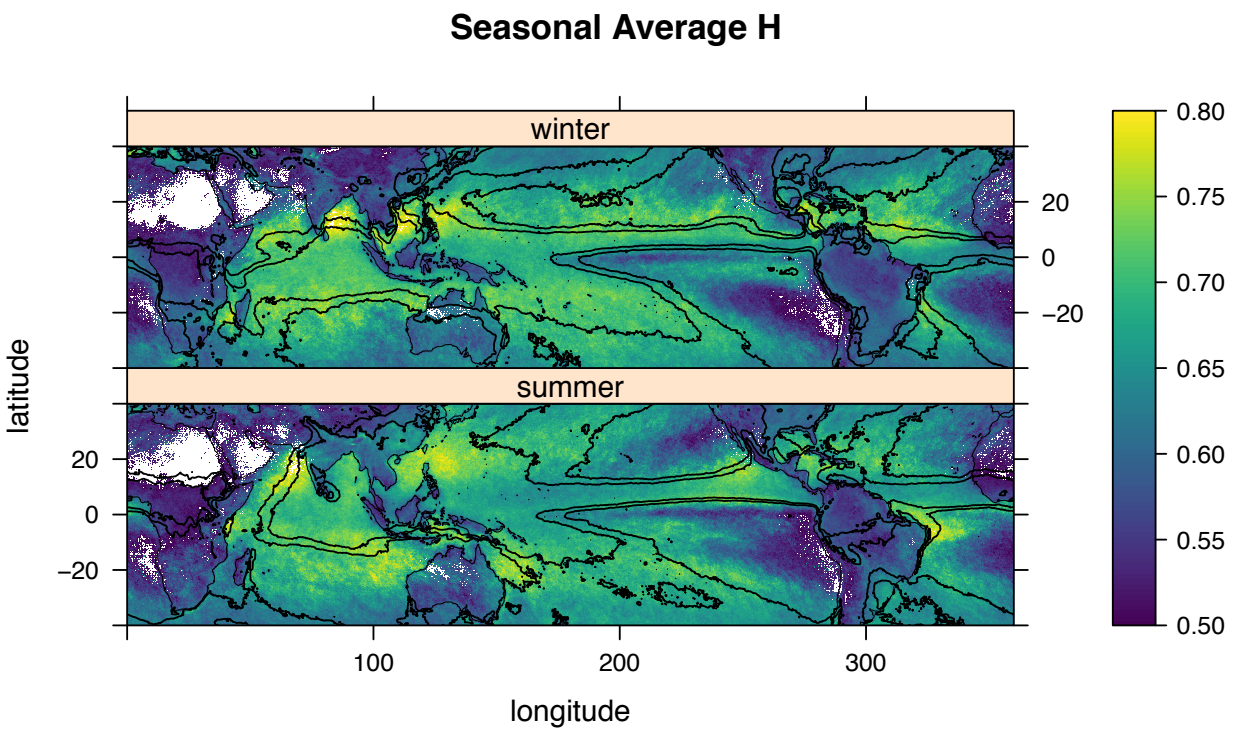

Figure 3: *Maps of seasonal average H for Northern Hemispheric summer (May–October, top) and winter (November–the next April, bottom). Contours indicate 2 and 4 mm day^{-1} average precipitation. White indicates locations where H is not defined.*

It is possible that the scaling behavior measured on scales from 12 hours to 16 days is better defined in some locations than others. In particular, the scaling exponent H can actually take different values over different frequency bands; the transition point between scaling regimes is

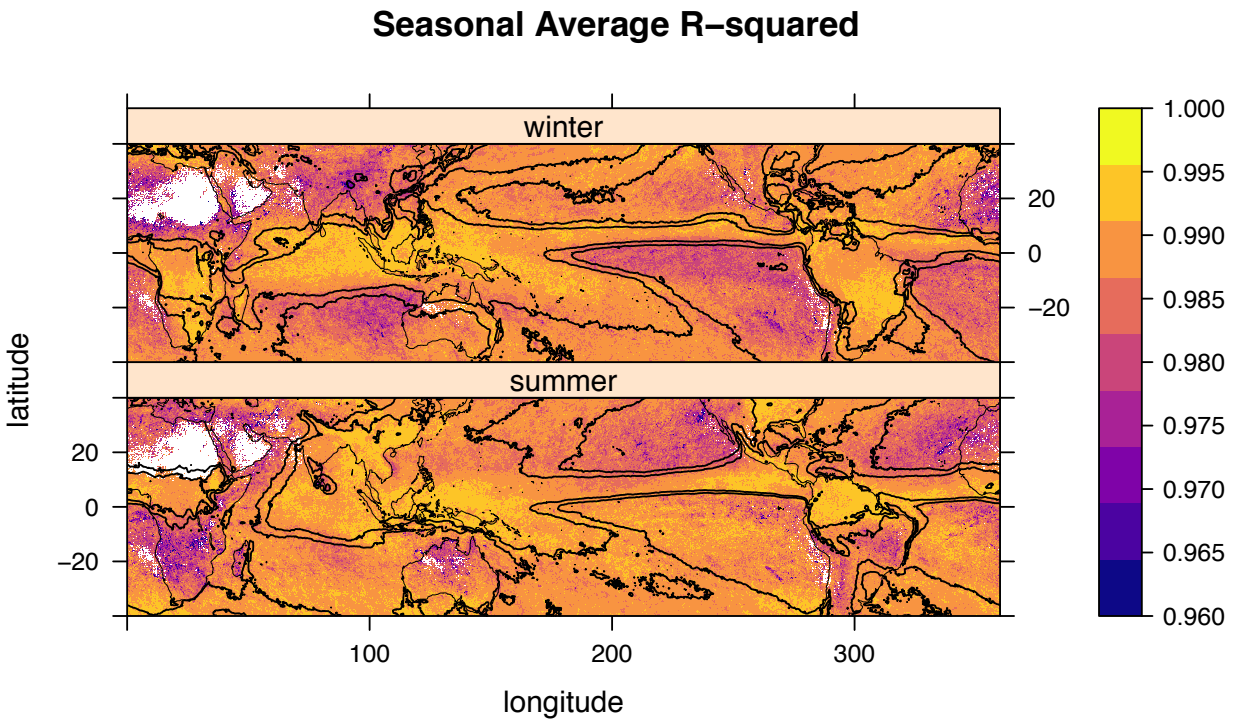

Figure 4: Maps of seasonal average R^2 obtained in fitting H. Contours indicate 2 and 4 mm day^{-1} average precipitation. White indicates locations where H is not defined.

called a "scaling break" (e.g., Gao et al., 2011) or "crossover" (e.g., Telesca et al., 2012). To obtain a heuristic indication of possible scaling breaks we use the coefficient of determination R^2 from the linear models used to estimate H. Low R^2 indicates that there is variability in $\log_2 F(w)$ that is not accounted for by the single linear fit, i.e., a scaling break.

Figure 4 shows the seasonal average maps of R^2. The seasonal average R^2 are obtained analogously to the mean H in Fig. 3, i.e., they are the average R^2 of either 17 winter subsets or 16 summer subsets for each location. In general, the fit quality indicated by R^2 tends to be very high in regions with high seasonal average precipitation as well as over land, especially in the summer hemisphere. There is some tendency for regions with low mean precipitation to have lower quality fits. This could be indicative of the presence of scaling breaks in the band from 12 hours to 16 days. It could also reflect increased sampling variability in the fluctuation function $F(w)$ due to the smaller number of positive precipitation rate observations available in certain arid locations. Further diagnosis of the additional scaling regimes that could be present in locations with lower quality fits is an excellent subject for future investigations.

4.6 D&R Meeting Data Analysis Challenges

Phenomena in Earth's weather and climate systems are inherently complex, largely owing to nonlinear interactions among its physical components. One classical example of such complexity is the multiscale, or "hierarchical", structure of tropical atmospheric convection (Nakazawa, 1988): con-

vective organization over the tropical Indo-Pacific sector appears to have spatial sizes from $O(10^2)$ to $O(10^3)$ km and temporal scales from $O(1)$ to $O(10)$ days. Fractal and multifractal concepts originally developed to understand turbulence phenomena (e.g., Frisch, 1995; Barenblatt, 1996) have been applied to phenomenologically describe the multiscale atmospheric convection and related physical processes, such as the persistence of tropical precipitation observed in Fig. 3. Despite of the parsimonious mathematical expression, conventional numerical weather and climate models that are among the essential tools for solving complex global issues have difficulty simulating or forecasting the scaling behavior in Eq.(2).

Bjerknes (1904) had dispelled a strictly mathematically analytical solution for the weather forecast problem for its over-simplification of multiscale problems. Instead, Bjerknes prophetically, perhaps in a whim, depicted "graphical or mixed graphical and numerical methods" that evolve a map based on initial observations to that representing the new state of the atmosphere, and remarked that "everything will depend on succeeding in breaking down the problem, which is overwhelmingly difficult as a whole, into partial problems of which none represents insurmountable difficulties".

More than a century has passed. The progress of finding and articulating the laws of the evolving multiscale atmospheric dynamics and physics remains excruciatingly slow. Focused on atmospheric convection, Arakawa et al. (2011, 2016) proposed two routes to unify the multiscale modeling of the atmosphere. The first route is a unified convection scheme applicable to a range of horizontal mesh resolutions between coarse ones typically used by global climate models (GCMs) and fine cloud-resolving models (CRMs); the second route is a multiscale-modeling framework that implements explicit representation of convection and associated cloud-scale processes by CRMs. These theoretical and the ensuing modeling developments have not been matched with a data analysis framework for validation and uncertainty quantification. To gain insights from the immense and complex model outputs along with observational data, deep analysis with D&R is one promising approach.

5 Summary and Conclusions

The Divide & Recombine statistical approach to big data and high computational complexity of analytic methods enables deep analysis, that is, analysis of data at their finest granularity including visualization (Sec. 2). It is tempting when the data are big to first carry out dimension reduction or to analyze just summary statistics. This greatly increases the risk of missing important information in the data. Deep analysis is possible with small data and is widely practiced. D&R extends this to big data.

The **DeltaRho** D&R software provides high-performance parallel, distributed computing for the analysis of big data and high computational complexity of analytic methods. It has the **Hadoop** distributed parallel compute engine and database at the back end. An analyst programs analyses at the front end with **R** using the **DeltaRhoR**-package **datadr**, a domain-specific language for D&R. This makes programming D&R analyses very efficient for the data analyst. In the lifecycle of a big-data analysis project, more time can be spent on thinking about the data analyzed and the subject matter, and less time programming. In addition, **datadr** can be run by itself on a multicore server and manage parallel computing itself with data in memory or on disk.

With **Hadoop** as the back end, we illustrated D&R by application to two big, complex datasets that benefitted greatly from D&R with **DeltaRho**. One is for the Spamhaus IP address blacklisting service (Sec. 3). The data are values of 13 variables for 10,615,054,608 queries of the status of IP addresses. We briefly discuss analysis of the profiles of 59,435,635 unique queried IP addresses with at least one blacklisting. The profiles are marked point processes whose number of points, a very skewed distribution, ranges from 16 to 2,097,152. The marks for each are values of 8 variables. Deep analysis will be required to understand and model these processes.

The other is the Tropical Rainfall Measuring Mission (TRMM) data (Sec. 4). Our study of the TRMM data illuminates the persistent temporal clustering patterns of tropical rainfall through their correlation structure as quantified by the Hurst parameter H. We estimated the H for 12-hr to 16-day time-scale of interests to operational atmospheric scientists, conditioned on seasons in the global Tropics with $\sim$ 17-years of 3-hourly TRMM data. We pointed out that while two regions may have very similar seasonal average precipitation, one may have very consistent rainfall while another has an envelope of extreme rainfall events followed by drought. The Hurst parameter conveniently discriminates between these patterns. Furthermore, we showed that the seasonal average clustering patterns feature a land-sea contrast with much stronger clustering over the oceans. High temporal clustering tends to occur in regions with strong horizontal gradient of the mean precipitation and near coastal boundaries. These patterns are of great practical importance for water resource management since they control how the local seasonal water supply becomes available over time.

6 Acknowledgements

D&R and **DeltaRho** were supported by the NSF/DHS Visual Analytics Program Award 0937123, the NSF CDS&E Big Data Program Award 1228348, and the DARPA XDATA Big Data Program Contract FA8750-12-2-0343. WT and MCB were partially supported by the NASA Earth and Space Science Fellowship grant NASA-NNX16AO62H. The authors are grateful to Doug Crabill for helping maintain the **DeltaRho** software stack and administrating the **Hadoop** clusters. They thank Qi Liu for assisting with TRMM data ingestion. This research was supported in part through computational resources provided by Information Technology at Purdue, West Lafayette, Indiana.

References

A. Arakawa, J.-H. Jung, and C.-M. Wu. Toward unification of the multiscale modeling of the atmosphere. *Atmos. Chem. Phys.*, 11(8):3731–3742, apr 2011. ISSN 1680-7324. doi: 10.5194/acp-11-3731-2011.

Akio Arakawa, Joon-Hee Jung, and Chien-Ming Wu. Multiscale Modeling of the Moist-Convective Atmosphere. In Robert G. Fovell and Wen-wen Tung, editors, *Meteorol. Monogr.*, volume 56, pages 16.1–16.17. American Meteorological Society, 2016. doi: 10.1175/AMSMONOGRAPHS-D-15-0014.1.

G. I. Barenblatt. *Scaling, Self-Similarity and Intermediate Asymptotics*. Cambridge University Press, 1996. ISBN 9780521435222.

Vilhelm Bjerknes. Das Problem der Wettervorhersage, betrachtet vom Standpunkte der Mechanik und der Physik (The problem of weather prediction, considered from the viewpoints of mechanics and physics, trans. and ed. by E. Volken and S. Brönmann, Meteorol. Z. 18 (2009)). *Meteorol. Z.*, 21:1–7, 1904.

M. C. Bowers, J. B. Gao, and W.-w. Tung. Long range correlations in tree ring chronologies of the USA: Variation within and across species. *Geophys. Res. Lett.*, 40(September 2012):1–5, feb 2013. ISSN 00948276. doi: 10.1029/2012GL054011.

D. R Brillinger. John W. Tukey's work on time series and spectrum analysis. *Ann. Stat.*, 30(6): 1595–1618, 2002. doi: 10.1214/aos/1043351248.

William S. Cleveland. Data Science: An Action Plan for Expanding the Technical Areas of. *Int. Stat. Rev.*, 4(5):497–511, 2001. ISSN 19321872. doi: 10.1002/sam.

William S. Cleveland. Learning from Data : Unifying Statistics and Computer Science. *Int. Stat. Rev.*, 73(2):217–221, 2005. ISSN 03067734, 17515823.

William S. Cleveland and Susan J. Devlin. Locally weighted regression: An approach to regression analysis by local fitting. *J. Am. Stat. Assoc.*, 83(403):596–610, 1988. ISSN 1537274X. doi: 10.1080/01621459.1988.10478639.

William S. Cleveland and Ryan Hafen. Divide and recombine (D&R): Data science for large complex data. *Stat. Anal. Data Min.*, 7(6):425–433, 2014. ISSN 19321872. doi: 10.1002/sam.11242.

Christopher Davis, Barbara Brown, and Randy Bullock. Object-Based Verification of Precipitation Forecasts. Part I: Methodology and Application to Mesoscale Rain Areas. *Mon. Weather Rev.*, 134(7):1772–1784, 2006a. ISSN 0027-0644. doi: 10.1175/MWR3145.1.

Christopher Davis, Barbara Brown, and Randy Bullock. Object-Based Verification of Precipitation Forecasts. Part II: Application to Convective Rain Systems. *Mon. Weather Rev.*, 134(7):1785–1795, 2006b. ISSN 0027-0644. doi: 10.1175/MWR3146.1.

Uriel Frisch. *Turbulence: The Legacy of A. N. Kolmogorov*. Cambridge University Press, 1995. ISBN 0521457130.

J. B. Gao, Y. Cao, W.-w. Tung, and J. Hu. *Multiscale Analysis of Complex Time Series: Integration of Chaos and Random Fractal Theory, and Beyond*. John Wiley & Sons, Inc., 2007. ISBN 9780471654704. doi: 10.1002/9780470191651.

Jianbo Gao, Jing Hu, Wen-Wen Tung, Yinhe Cao, N. Sarshar, and Vwani Roychowdhury. Assessment of long-range correlation in time series: How to avoid pitfalls. *Phys. Rev. E*, 73(1):1–10, jan 2006. ISSN 1539-3755. doi: 10.1103/PhysRevE.73.016117.

Jianbo Gao, Jing Hu, and Wen-wen Tung. Facilitating joint chaos and fractal analysis of biosignals through nonlinear adaptive filtering. *PLoS One*, 6(9):e24331, jan 2011. ISSN 1932-6203. doi: 10.1371/journal.pone.0024331.

Saptarshi Guha, Paul Kidwell, Ryan P. Hafen, and William S. Cleveland. Visualization Databases for the Analysis of Large Complex Datasets. *J. Mach. Learn. Res.*, 5:193–200, 2009. ISSN 15324435.

Saptarshi Guha, Ryan Hafen, Jeremiah Rounds, Jin Xia, Jianfu Li, Bowei Xi, and William S. Cleveland. Large complex data: Divide and recombine (D&R) with RHIPE. *Stat*, 1(1):53–67, 2012. ISSN 20491573. doi: 10.1002/sta4.7.

R. P. Hafen, W.S. Cleveland, and L. H. Sego. DeltaRho, www.deltarho.org, 2016.

George J. Huffman, David T. Bolvin, Eric J. Nelkin, David B. Wolff, Robert F. Adler, Guojun Gu, Yang Hong, Kenneth P. Bowman, and Erich F. Stocker. The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales. *J. Hydrometeorol.*, 8(1):38–55, feb 2007. ISSN 1525-755X. doi: 10.1175/JHM560.1.

H. E. Hurst. Long-term storage capacity of reservoirs. *Trans. Am. Soc. Civ. Eng.*, 116(1):770–799, 1951.

S. Lovejoy and B. B. Mandelbrot. Fractal properties of rain, and a fractal model. *Tellus A*, 37 A (3):209–232, 1985. ISSN 16000870. doi: 10.1111/j.1600-0870.1985.tb00423.x.

Helena Mitasova, Russell S. Harmon, Katherine J. Weaver, Nathan J. Lyons, and Margery F. Overton. Scientific visualization of landscapes and landforms. *Geomorphology*, 137(1):122–137, 2012. ISSN 0169555X. doi: 10.1016/j.geomorph.2010.09.033.

Tetsuo Nakazawa. Tropical Super Clusters within Intraseasonal Variations over the Western Pacific. *J. Meteorol. Soc. Japan*, 66(6):823–839, 1988.

C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger. Mosaic Organization of DNA Nucleotides. *Phys. Rev. E*, 49(2):1685–1689, 1994.

Scott Sellars, Phu Nguyen, Wei Chu, Xiaogang Gao, Kuo-lin Hsu, and Soroosh Sorooshian. Computational Earth Science: Big Data Transformed Into Insight. *EOS, Trans. Am. Geophys. Union*, 94(32):277–278, 2013. ISSN 2324-9250. doi: 10.1002/2013EO320001.

Scott Lee Sellars, Xiaogang Gao, and Soroosh Sorooshian. An Object-Oriented Approach to Investigate Impacts of Climate Oscillations on Precipitation: A Western United States Case Study. *J. Hydrometeorol.*, 16:830–842, 2015. ISSN 1525-755X. doi: 10.1175/JHM-D-14-0101.1.

J. Shi, Y. Qiu, U.F. Minhas, L. Jiao, C. Wang, B. Reinwald, and F. Özcan. Clash of the titans: Mapreduce vs. spark for large scale data analytics. *Proc. VLDB Endow.*, 8(13):2110–2121, 2015. ISSN 21508097 (ISSN). doi: 10.14778/2831360.2831365.

J. Simpson, C. Kummerow, W. K. Tao, and R. F. Adler. On the Tropical Rainfall Measuring Mission (TRMM). *Meteorol. Atmos. Phys.*, 60(1-3):19–36, 1996. ISSN 0177-7971. doi: 10.1007/BF01029783.

Soroosh Sorooshian, Kuo Lin Hsu, Xiaogang Gao, Hoshin V. Gupta, Bisher Imam, and Dan Braithwaite. Evaluation of PERSIANN system satellite-based estimates of tropical rainfall. *Bull. Am. Meteorol. Soc.*, 81(9):2035–2046, 2000. ISSN 00030007. doi: 10.1175/1520-0477(2000)081⟨2035:EOPSSE⟩2.3.CO;2.

L Telesca, J Pierini, and B Scian. Investigating the temporal variation of the scaling behavior in rainfall data measured in central Argentina by means of detrended fluctuation analysis. *Phys. A Stat. Mech. its Appl.*, 391(4):1553–1562, 2012. doi: 10.1016/J.PHYSA.2011.08.042.

Wen-wen Tung, Dimitrios Giannakis, and Andrew J. Majda. Symmetric and Antisymmetric Convection Signals in the Madden-Julian Oscillation. Part I: Basic Modes in Infrared Brightness Temperature. *J. Atmos. Sci.*, 71(9):3302–3326, 2014. ISSN 0022-4928. doi: 10.1175/JAS-D-13-0122.1.

Michelle T. H. van Vliet, David Wiberg, Sylvain Leduc, and Keywan Riahi. Power-generation system vulnerability and adaptation to changes in climate and water resources. *Nat. Clim. Chang.*, 6:375–380, 2016. ISSN 1758-678X. doi: 10.1038/nclimate2903.

John K. Williams. Using random forests to diagnose aviation turbulence. *Mach. Learn.*, 95(1):51–70, 2014. ISSN 15730565. doi: 10.1007/s10994-013-5346-7.

World Economic Forum. *The Global Risks Report 2017 12th Edition*. World Economic Forum, Geneva, 2017. ISBN 080116. doi: 10.1017/CBO9781107415324.004.

Tetsuzo Yasunari. The Monsoon Year—A New Concept of the Climatic Year in the Tropics. *Bull. Am. Meteorol. Soc.*, 72(9):1331–1338, 1991. ISSN 0003-0007. doi: 10.1175/1520-0477(1991)072⟨1331:TMYNCO⟩2.0.CO;2.