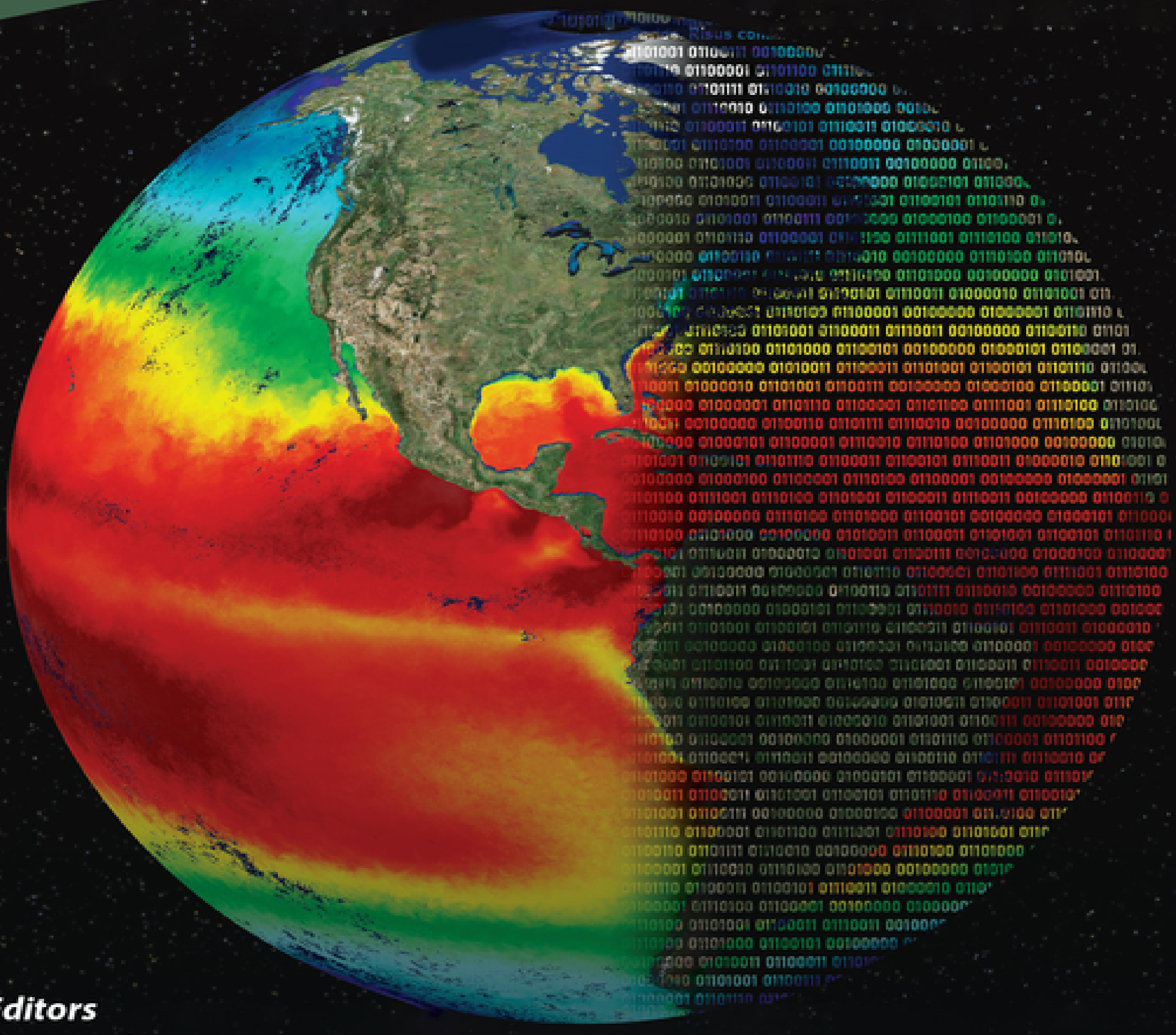
SPECIAL PUBLICATIONS SERIES

AGU
ADVANCING EARTH
AND SPACE SCIENCE

Big Data Analytics in Earth, Atmospheric, and Ocean Sciences



Editors
Thomas Huang
Tiffany C. Vance
Christopher Lynnes

WILEY

Table of Contents

List of Tables

List of Illustrations

Chapter 14

Chapter 15

Special Publications 77

BIG DATA ANALYTICS IN EARTH, ATMOSPHERIC, AND OCEAN SCIENCES

Thomas Huang
Tiffany C. Vance
Christopher Lynnes

Editors

This Work is a co-publication of the American Geophysical Union and John Wiley and Sons, Inc.

This Work is a co-publication of the American Geophysical Union and John Wiley and Sons, Inc.

This edition first published 2023

Published under the aegis of the AGU Publications Committee

Matthew Giampoala, Vice President, Publications

Carol Frost, Chair, Publications Committee

For details about the American Geophysical Union visit us at www.agu.org.

Registered Office

John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office

111 River Street, Hoboken, NJ 07030, USA

For details of our global editorial offices, customer services, and more information about Wiley products visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data

Names: Huang, Thomas (Technologist), editor. | Vance, C. Tiffany, editor. | Lynnes, Christopher, editor.

Title: Big data analytics in earth, atmospheric, and ocean sciences / Thomas Huang, Vance, C. Tiffany, Christopher Lynnes, editors.

Description: Hoboken, NJ : Wiley-American Geophysical Union, 2023. | Includes bibliographical references and index.

Identifiers: LCCN 2022020168 (print) | LCCN 2022020169 (ebook) | ISBN 9781119467571 (cloth) | ISBN 9781119467564 (adobe pdf) | ISBN 9781119467533 (epub)

Subjects: LCSH: Earth sciences–Data processing. | Atmospheric science–Data processing. | Marine sciences–Data processing. | Big data.

Classification: LCC QE48.8 .B54 2022 (print) | LCC QE48.8 (ebook) | DDC 550.0285/57–dc23/eng20220722

LC record available at https://lccn.loc.gov/2022020168

LC ebook record available at https://lccn.loc.gov/2022020169

Cover Design: Wiley

Cover Images: Courtesy of Kate Culpepper with design elements provided by Esri, HERE, Garmin, FAO, NOAA, USGS, EPA | Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community

List of Contributors

Edward M. Armstrong
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Alberto Arribas
Microsoft
Reading, UK; and
Department of Meteorology
University of Reading
Reading, UK

Jessica Austin
Axiom Data Science, LLC
Anchorage, Alaska, USA

Rob Bochenek
Axiom Data Science, LLC
Anchorage, Alaska, USA

Mark A. Bourassa
Center for Ocean-Atmospheric Prediction Studies, and
Department of Earth, Ocean, and Atmospheric Science
Florida State University
Tallahassee, Florida, USA

Jonathan Brannock

North Carolina Institute for Climate Studies

NOAA Cooperative Institute for Satellite Earth System Studies

North Carolina State University

Asheville, North Carolina, USA

Otis Brown

North Carolina Institute for Climate Studies

NOAA Cooperative Institute for Satellite Earth System Studies

North Carolina State University

Asheville, North Carolina, USA

Kevin A. Butler

Environmental Systems Research Institute

Redlands, California, USA

Nga T. Chung

NASA Jet Propulsion Laboratory

California Institute of Technology

Pasadena, California, USA

Thomas Cram

National Center for Atmospheric Research

Boulder, Colorado, USA

Jenny Dissen

North Carolina Institute for Climate Studies

NOAA Cooperative Institute for Satellite Earth System Studies

North Carolina State University

Asheville, North Carolina, USA

Kyle Dumas

ARM Research Facility

Oak Ridge National Laboratory

Oak Ridge, Tennessee, USA

John-Marc Dunaway

Axiom Data Science, LLC

Anchorage, Alaska, USA

Jocelyn Elya

Center for Ocean-Atmospheric Prediction Studies

Florida State University

Tallahassee, Florida, USA

Eamon Ford

NASA Jet Propulsion Laboratory

California Institute of Technology

Pasadena, California, USA

David W. Gallaher

National Snow and Ice Data Center

Boulder, Colorado, USA

Kevin Michael Gill

NASA Jet Propulsion Laboratory

California Institute of Technology
Pasadena, California, USA

Glenn E. Grant
National Snow and Ice Data Center
Boulder, Colorado, USA

Frank R. Greguska III
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Mahabaleshwara Hegde
NASA Goddard Space Flight Center
Greenbelt, Maryland, USA

Alex Held
CSIRO Centre for Earth Observation
Canberra, ACT, Australia

Erik Hoel
Environmental Systems Research Institute
Redlands, California, USA

Benjamin Holt
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Hook Hua

NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Thomas Huang
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Joseph C. Jacob
NASA Jet Propulsion Laboratory
Pasadena, California, USA

Zaihua Ji
National Center for Atmospheric Research
Boulder, Colorado, USA

Karthik Kashinath
Lawrence Berkeley National Laboratory
Berkeley, California, USA

Edward J. Kearns
First Street Foundation
Brooklyn, New York, USA

Bhargavi Krishna
ARM Research Facility
Oak Ridge National Laboratory
Oak Ridge, Tennessee, USA

Vitaliy Kurlin

Department of Computer Science

University of Liverpool

Liverpool, UK

Michael M. Little

NASA Goddard Space Flight Center

Greenbelt, Maryland, USA

Qin Lv

Department of Computer Science

University of Colorado

Boulder, Colorado, USA

Christopher Lynnes

NASA Goddard Space Flight Center (retd.)

Greenbelt, Maryland, USA

Gerald Manipon

NASA Jet Propulsion Laboratory

California Institute of Technology

Pasadena, California, USA

Theo McCaie

Met Office

Exeter, UK

Dmitriy Morozov

Lawrence Berkeley National Laboratory

Berkeley, California, USA

Grzegorz Muszynski

Lawrence Berkeley National Laboratory

Berkeley, California, USA; and

Department of Computer Science

University of Liverpool

Liverpool, UK

Matt Paget

CSIRO Centre for Earth Observation

Canberra, ACT, Australia

Tom Powell

Met Office

Exeter, UK

Giri Prakash

ARM Research Facility

Oak Ridge National Laboratory

Oak Ridge, Tennessee, USA

Prabhat Ram

Lawrence Berkeley National Laboratory

Berkeley, California, USA

Niall Robinson

Met Office

Exeter, UK; and

University of Exeter
Exeter, UK

Sujen Shah
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Adrienne Simonson
Office of the Chief Information Officer
National Oceanic and Atmospheric Administration
Asheville, North Carolina, USA

Shawn R. Smith
Center for Ocean-Atmospheric Prediction Studies
Florida State University
Tallahassee, Florida, USA

Kate Szura
Interactions LLC
Franklin, Massachusetts, USA

Ronnie Taib
CSIRO Data61
Sydney, NSW, Australia

Jacob Tomlinson
NVIDIA
Reading, UK

Vardis Tsontos
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Tiffany C. Vance
U.S. Integrated Ocean Observing System
National Oceanic and Atmospheric Administration
Silver Spring, Maryland, USA

Peter Wang
CSIRO Data61
Sydney, NSW, Australia

Michael Wehner
Lawrence Berkeley National Laboratory
Berkeley, California, USA

Brian D. Wilson
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Robert Woodcock
CSIRO Centre for Earth Observation
Canberra, ACT, Australia

Elizabeth Yam
NASA Jet Propulsion Laboratory

California Institute of Technology
Pasadena, California, USA

Chaowei Phil Yang
George Mason University
Fairfax, Virginia, USA

Alice Yepremyan
NASA Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California, USA

Hailiang Zhang
NASA Goddard Space Flight Center
Greenbelt, Maryland, USA

Preface

The seeds for this book were sown in sessions on Big Data Analytics, held at the 2016 Fall Meeting of the American Geophysical Union. At the time, Earth Science data were projected to rise by orders of magnitude in the coming decade, and the community was investigating a variety of emergent technologies and techniques to make the best use of the coming deluge. The chapters of this book are a representative, but by no means exhaustive, collection of those and similar investigations.

Big Earth Data Analytics can be defined as the application of increasingly sophisticated tools for data analysis and display to the rapidly increasing volume of Earth science data to obtain information, and eventually insight. This combines two concepts: Big Earth Data and Data Analytics. Big Earth Data refers both to the volume of data sets and the combination of data from a variety of sources, in a variety of formats, and from a variety of disciplines. To get a sense of the volume, NOAA generates tens of terabytes of data a day from satellites, radars, ships, weather models, and other sources. The National Aeronautics and Space Administration (NASA) Earth Observation archives were growing by more than 30 TB per day in 2020 with daily growth expected to increase to 130 TB/day by 2024 as new satellites launch; and the European Centre for Medium-Range Weather Forecasts (ECMWF) meteorological data archive adds 200 terabytes of new data daily. However, the data are "big" not only in their volume but in their varied formats, disciplines, structures, and formats. As such, they are disruptors to traditional analysis methods, and to the kinds of questions that can be asked by researchers. Data analytics are increasingly driven by the availability of high-volume and heterogeneous data sets. Data size and complexity affect all aspects of data management and usage, requiring new approaches and tools. Despite the challenges to acquire, use, and analyze Big Earth Data, they are already being utilized extensively in climate, oceanographic, and biology related works. Easily available data lead to the ability to analyze longer scale records and patterns over large spatial domains.

Analyses of these data borrow both from traditional scientific analyses and from tools developed for business applications. These types of data analytics are developed by university and other research teams. They are increasingly becoming an area of interest to cloud providers and analytics companies. From Google's Earth Engine for analyzing Earth science data at scale, to the National Oceanic and Atmospheric Administration's (NOAA's) Big Data Program, big data about the Earth and their analysis are increasingly common. Amazon's Elastic MapReduce and SageMaker are common building blocks for cloud-based analysis and Galileo (a.k.a. Service Workbench) is Amazon's latest Web application for interactive analysis. Microsoft Azure ML Studio is another popular cloud-based data analysis solution. Big Earth Data analyses increasingly rely on cloud-based storage and processing capabilities as the volume of the data and the computing resources needed go beyond local resources.

This book is organized into three parts. It starts with the big picture, covering Big Data Analytics Architecture. This part begins with a chapter addressing the geospatial aspect of Big Earth Data from a variety of perspectives. This is followed by a chapter discussing the data management challenges posed by data at scale, particularly in the context of making them available for analysis. This is complemented by a chapter discussing the challenges of scaling up the analysis itself. The following chapters cover large-scale projects such as NASA's Earth Exchange, which enables large scale data analysis in a supercomputing environment and the NOAA Big Data Project, which makes data sets available to end users via several cloud providers. Part I also includes chapters on architectures and fully realized systems, such as Data Cube, NEXUS and the Apache Science Data Analytics Platform, and a NoSQL based platform for exploring and analyzing in situ data.

The second part of the book, Analysis Methods for Big Earth Data, addresses some specific techniques to derive information and/or insight from big data, emphasizing the unique aspects of Earth Observations. Part II begins with two chapters on the use of geospatial statistics for analysis, followed by a chapter melding machine learning with geophysical constraints, and finally a chapter benchmarking different analytical methods for spatiotemporal analysis.

The third part of the book, Big Earth Data Applications, describes a few specific applications of big analysis techniques and platforms: weather and climate model analysis, atmospheric river patterns, Antarctic land surface temperatures extremes, satellite in situ match-ups of oceanographic data, and vessel tracking. This is clearly a small sample of existing applications; rather, the sample shows how some very different analysis methods can find diverse applications in the Earth sciences.

While the application of big Earth data analytics covers a range of applications, a number of common themes in the chapters of this book include (1) the role of the cloud, especially with ever increasing data sizes; (2) limitations and costs of using the cloud, including the unpredictability of costs and the high cost of data egress from the cloud; (3) techniques to maintain data integrity during file transfers; (4) efficiencies via partial reads from Web object storage;