STATISTICAL METHODS FOR CLIMATE SCIENTISTS

This book provides a comprehensive introduction to the most commonly used statistical methods relevant in atmospheric, oceanic, and climate sciences. Each method is described step-by-step using plain language, and illustrated with concrete examples, with relevant statistical and scientific concepts explained as needed. Particular attention is paid to nuances and pitfalls, with sufficient detail to enable the reader to write relevant code. Topics covered include hypothesis testing, time series analysis, linear regression, data assimilation, extreme value analysis, Principal Component Analysis, Canonical Correlation Analysis, Predictable Component Analysis, and Covariance Discriminant Analysis. The specific statistical challenges that arise in climate applications are also discussed, including model selection problems associated with Canonical Correlation Analysis, Predictable Component Analysis, and Covariance Discriminant Analysis. Requiring no previous background in statistics, this is a highly accessible textbook and reference for students and early career researchers in the climate sciences.

TIMOTHY M. DELSOLE is Professor in the Department of Atmospheric, Oceanic and Earth Sciences, and Senior Scientist at the Center for Oceanic Atmospheric, and Land Studies, at George Mason University, Virginia. He has published more than 100 peer-reviewed papers in climate science and served as co-editor-in-chief of the *Journal of Climate*.

MICHAEL K. TIPPETT is an associate professor at Columbia University. His research includes forecasting El Niño and relating extreme weather (tornadoes and hurricanes) with climate, now and in the future. He analyzes data from computer models and weather observations to find patterns that improve understanding, facilitate prediction, and help manage risk.

Includes both the mathematics and the intuition needed for climate data analysis.

–Professor Dennis L Hartmann, *University of Washington*

STATISTICAL METHODS FOR CLIMATE SCIENTISTS

TIMOTHY M. DELSOLE
George Mason University

MICHAEL K. TIPPETT
Columbia University

CAMBRIDGE UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314-321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781108472418
DOI: 10.1017/9781108659055

First published 2022

Printed in the United Kingdom by TJ Books Limited, Padstow Cornwall

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Names: DelSole, Timothy M., author.
Title: Statistical methods for climate scientists / Timothy M. DelSole and Michael K. Tippett.
Description: New York : Cambridge University Press, 2021. | Includes bibliographical references and index.
Identifiers: LCCN 2021024712 (print) | LCCN 2021024713 (ebook) | ISBN 9781108472418 (hardback) | ISBN 9781108659055 (epub)
Subjects: LCSH: Climatology–Statistical methods. | Atmospheric science–Statistical methods. | Marine sciences–Statistical methods. | BISAC: SCIENCE / Earth Sciences / Meteorology & Climatology
Classification: LCC QC866 .D38 2021 (print) | LCC QC866 (ebook) | DDC 551.601/5118–dc23
LC record available at https://lccn.loc.gov/2021024712
LC ebook record available at https://lccn.loc.gov/2021024713

ISBN 978-1-108-47241-8 Hardback

Additional resources for this publication at www.cambridge.org/9781108472418.

Contents

Preface

This book provides an introduction to the most commonly used statistical methods in atmospheric, oceanic, and climate sciences. The material in this book assumes no background in statistical methods and can be understood by students with only a semester of calculus and physics. Also, no advanced knowledge about atmospheric, oceanic, and climate sciences is presumed. Most chapters are self-contained and explain relevant statistical and scientific concepts as needed. A familiarity with calculus is presumed, but the student need not solve calculus problems to perform the statistical analyses covered in this book.

The need for this book became clear several years ago when one of us joined a journal club to read "classic" papers in climate science. Specifically, students in the club had difficulty understanding certain papers because these papers contained unfamiliar statistical concepts, such as empirical orthogonal functions (EOFs), significance tests, and power spectra. It became clear that our PhD curriculum was not adequately preparing students to be "literate" in climate science. To rectify this situation, we decided that students should take a statistics class. However, at that time, there did not exist a single self-contained course that covered all the topics that we considered to be essential for success in climate science. Therefore, we designed a single course that covered these topics (which eventually expanded into a two-semester course). This book is based on this course and embodies over a decade of experience in teaching this material.

This book covers six key statistical methods that are essential to understanding modern climate research: (1) hypothesis testing; (2) time series models and power spectra; (3) linear regression; (4) Principal Component Analysis (PCA), and related multivariate decomposition methods such as Canonical Correlation Analysis (CCA) and Predictable Component Analysis, (5) data assimilation; and (6) extreme value analysis. Chapter 1 reviews basic probabilistic concepts that are used throughout the book. Chapter 2 discusses hypothesis testing. Although the likelihood ratio provides a general framework for hypothesis testing, beginners often find this framework

daunting. Accordingly, Chapter 2 explains hypothesis testing based on heuristic arguments for Gaussian distributions, which most students find intuitive. The framework discussed in Chapter 2 provides the foundation for hypothesis testing that is used in the rest of the book. The related concept of confidence intervals, as well as bootstrap methods and distribution-free tests, is discussed in Chapters 3 and 4. Fundamental concepts in time series analysis, especially stochastic processes and power spectra, are discussed in Chapters 5 and 6, respectively. Certain topics that typically are included in statistical texts are omitted because they are seldom used in climate science; for instance, moving average models are not discussed in detail because they are used much less often in climate science than autoregressive models.

The second half of this book covers multivariate methods. We have striven to convey our hard-learned experience about these methods collected over many years. Basic concepts in linear algebra and multivariate distributions are outlined in Chapter 7. Linear regression is discussed in Chapters 8 and 9. Pitfalls in linear regression are discussed in detail, especially model selection (Chapter 10) and screening (Chapter 11). These concepts are critical for proper usage and interpretation of statistical methods, especially in statistical prediction, but are not easy to find in introductory texts. Principal Component Analysis is the most commonly used multivariate method in climate science, hence our discussion in Chapter 12 is very detailed. Subsequent chapters discuss field significance (Chapter 13), Multivariate Linear Regression (Chapter 14), Canonical Correlation Analysis (Chapter 15), Covariance Discriminant Analysis (Chapter 16), Analysis of Variance (Chapter 17), and Predictable Component Analysis (Chapter 18). An introduction to extreme value theory is provided in Chapter 19. Data assimilation and ensemble square root filters are discussed in Chapters 20 and 21 with the goal of introducing essential ideas and common practical problems that we believe every user of data assimilation products should be aware of.

This book is designed for either a one-semester or a two-semester course. Considerable effort has been made to select and arrange the material in a logical order that facilitates teaching and learning. We have used this book to teach a one-semester course covering Chapters 1–13 at approximately one chapter per week. For more advanced students, a second-semester course is offered covering Chapters 14–21. The homework sets are available at the Cambridge University Press website associated with this book.

The multivariate part of this book is distinguished from previous books in an important way. Typical climate data sets are much bigger in the spatial dimension than in the time dimension. This creates major difficulties for applying such multivariate techniques as Canonical Correlation Analysis, Predictable Component Analysis, and Covariance Discriminant Analysis to climate data, although these

difficulties are rarely discussed in standard statistics texts. In the climate literature, the standard approach to this problem is to apply these techniques to a few principal components of the data, so that the time dimension is much bigger than the state dimension. The most outstanding barrier in this approach is choosing the number of principal components. Unfortunately, no standard criterion for selecting the number of principal components exists for these multivariate techniques. This gap was sorely felt each time this material was taught and motivated us to conduct our own independent research into this problem. This research culminated in the discovery of a criterion that was consistent with standard information criteria and could be applied to all of the problems discussed in this book. For regression models and CCA, this criterion is called Mutual Information Criterion (MIC) and is introduced in Chapter 14 (for full details, see DelSole and Tippett, 2021a). After formulating this criterion, we discovered that it was consistent with many of the criteria derived by Fujikoshi et al. (2010) based on likelihood ratio methods, which supports the soundness of MIC. However, MIC is considerably easier to derive and apply. We believe that MIC will be of wide interest to statisticians and to scientists in other fields who use these multivariate methods.

The development of this book is somewhat unique. Initially, we followed our own personal experience by giving formal lectures on each chapter. Inspired by recent educational research, we began using a "flipped classroom" format, in which students read each chapter and sent questions and comments electronically *before* coming to class. The class itself was devoted to going over the questions/comments from students. We explicitly asked students to tell us where the text failed to help their understanding. To invite feedback, we told students that we needed their help in writing this book, because over the ten years that we have been teaching this topic, we have become accustomed to the concepts and could no longer see what is wrong with the text. The resulting response in the first year was more feedback than we had obtained in all the previous years combined. This approach not only revolutionized the way we teach this material but gave us concrete feedback about where precisely the text could be improved. With each subsequent year, we experimented with new material and, if it did not work, tried different ways. This textbook is the outcome of this process over many years, and we feel that it introduces statistical concepts much more clearly and in a more accessible manner than most other texts.

Each chapter begins with a brief description of a statistical method and a concrete problem to which it can be applied. This format allows a student to quickly ascertain if the statistical method is the one that is needed. Each problem was chosen after careful thought based on intrinsic interest, importance in real climate applications, and instructional value.

Each statistical method is discussed in enough detail to allow readers to write their own code to implement the method (except in one case, namely extreme value

theory, for which there exists easy-to-use software in R). The reason for giving this level of detail is to ensure that the material is complete, self-contained, and covers the nuances and points of confusion that arise in practice. Indeed, we, as active researchers, often feel that we do not adequately understand a statistical method unless we have written computer code to perform that method. Our experience is that students gain fundamental and long-lasting confidence by coding each method themselves. This sentiment was expressed in an end-of-year course evaluation, in which one of our students wrote, "Before this course, I had used someone else's program to compute an EOF, but I didn't really understand it. Having to write my own program really helped me understand this method."

The methods covered in this book share a common theme: to quantify and exploit dependencies between X and Y. Different methods arise because each method is tailored to a particular probability distribution or data format. Specifically, the methods depend on whether X and Y are scalar or vector, whether the values are categorical or continuous, whether the distributions are Gaussian or not, and whether one variable is held fixed for multiple realizations of the other. The most general method for quantifying X-Y dependencies for multivariate Gaussian distributions is Canonical Correlation Analysis. Special cases include univariate regression (scalar Y), field significance (scalar X), or correlation (scalar X and scalar Y). In climate studies, multiple realizations of Y for fixed X characterize *ensemble* data sets. The most general method for quantifying X-Y dependencies in ensemble data sets is Predictable Component Analysis (or equivalently, Multivariate Analysis of Variance). Special cases include Analysis of Variance (scalar Y), and the t-test (scalar X and scalar Y). Many of these techniques have non-Gaussian versions. Linear regression provides a framework for exploiting dependencies to predict one variable from the other. Autoregressive models and power spectra quantify dependencies across time. Data assimilation provides a framework for exploiting dependencies to infer Y given X while incorporating "prior knowledge" about Y. The techniques for the different cases, and the chapter in which they are discussed, are summarized in Table 0.1.

Table 0.1. *Summary of methods for quantifying dependencies between X and Y.*

Y	X	Statistic or Procedure	Chapter
Vector	Vector	Canonical Correlation Analysis	15
Scalar	Vector	Multiple regression	9
Vector	Scalar	Field significance	13
Scalar	Scalar	Scalar regression or correlation	1
Ensemble and vector	Categorical	Predictable Component Analysis	18
Ensemble and scalar	Categorical	Analysis of Variance	17
Ensemble and scalar	Two categories	t-test	2

It is a pleasure to acknowledge helpful comments from colleagues who graciously gave up some of their busy time to read selected chapters in this book, including Jeffrey Anderson, Grant Branstator, Ian Jolliffe, and Jagadish Shukla. We thank our (former) students whose feedback was invaluable to finding the best pedagogical approach to this material, especially Paul Buchman, Xia Feng, Rachael Gaal, Olivia Gozdz, Liwei Jia, Keri Kodama, Emerson LaJoie, Douglas Nedza, Abhishekh Srivastava, Xiaoqin Yan, and M. Tugrul Yilmaz. We are indebted to Anthony Barnston, Grant Branstator, Ping Chang, Ben Kirtman, Andy Majda, Tapio Schneider, Jagadish Shukla, and David Straus for discussions over many years that have shaped the material presented in this book. We thank Vera Akum for assistance in acquiring the permissions for the quotes that open each chapter. Special thanks to Tony Barnston for suggesting the example used in Chapter 9. Any errors or inaccuracies in this book rest solely with the authors. We will be grateful to readers who notify us of errors or suggestions for improvement of this book.