

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.

Cambridge University Press
978-1-108-47241-8 — Statistical Methods for Climate Scientists
Timothy DelSole , Michael Tippett
Frontmatter
[More Information](#)

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
978-1-108-47241-8 — Statistical Methods for Climate Scientists
Timothy DelSole, Michael Tippett
Frontmatter
[More Information](#)

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

© Cambridge University Press 2022

This publication is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without the written
permission of Cambridge University Press.

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://lcn.loc.gov/2021024712>

LC ebook record available at <https://lcn.loc.gov/2021024713>

ISBN 978-1-108-47241-8 Hardback

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

Cambridge University Press has no responsibility for the persistence or accuracy of
URLs for external or third-party internet websites referred to in this publication
and does not guarantee that any content on such websites is, or will remain,
accurate or appropriate.

Contents

<i>Preface</i>	<i>page</i> xiii
1 Basic Concepts in Probability and Statistics	1
1.1 Graphical Description of Data	2
1.2 Measures of Central Value: Mean, Median, and Mode	4
1.3 Measures of Variation: Percentile Ranges and Variance	6
1.4 Population versus a Sample	8
1.5 Elements of Probability Theory	8
1.6 Expectation	11
1.7 More Than One Random Variable	13
1.8 Independence	16
1.9 Estimating Population Quantities from Samples	18
1.10 Normal Distribution and Associated Theorems	20
1.11 Independence versus Zero Correlation	27
1.12 Further Topics	28
1.13 Conceptual Questions	29
2 Hypothesis Tests	30
2.1 The Problem	31
2.2 Introduction to Hypothesis Testing	33
2.3 Further Comments on the t -test	40
2.4 Examples of Hypothesis Tests	43
2.5 Summary of Common Significance Tests	49
2.6 Further Topics	50
2.7 Conceptual Questions	51
3 Confidence Intervals	52
3.1 The Problem	53

vi	<i>Contents</i>	
	3.2	Confidence Interval for a Difference in Means 53
	3.3	Interpretation of the Confidence Interval 55
	3.4	A Pitfall about Confidence Intervals 57
	3.5	Common Procedures for Confidence Intervals 57
	3.6	Bootstrap Confidence Intervals 64
	3.7	Further Topics 67
	3.8	Conceptual Questions 68
4		Statistical Tests Based on Ranks 69
	4.1	The Problem 70
	4.2	Exchangeability and Ranks 71
	4.3	The Wilcoxon Rank-Sum Test 73
	4.4	Stochastic Dominance 78
	4.5	Comparison with the t -test 79
	4.6	Kruskal–Wallis Test 81
	4.7	Test for Equality of Dispersions 83
	4.8	Rank Correlation 85
	4.9	Derivation of the Mean and Variance of the Rank Sum 88
	4.10	Further Topics 92
	4.11	Conceptual Questions 93
5		Introduction to Stochastic Processes 94
	5.1	The Problem 95
	5.2	Stochastic Processes 100
	5.3	Why Should I Care if My Data Are Serially Correlated? 105
	5.4	The First-Order Autoregressive Model 109
	5.5	The AR(2) Model 117
	5.6	Pitfalls in Interpreting ACFs 119
	5.7	Solutions of the AR(2) Model 121
	5.8	Further Topics 122
	5.9	Conceptual Questions 124
6		The Power Spectrum 126
	6.1	The Problem 127
	6.2	The Discrete Fourier Transform 129
	6.3	Parseval’s Identity 133
	6.4	The Periodogram 134
	6.5	The Power Spectrum 135
	6.6	Periodogram of Gaussian White Noise 138
	6.7	Impact of a Deterministic Periodic Component 139

Contents

vii

6.8	Estimation of the Power Spectrum	140
6.9	Presence of Trends and Jump Discontinuities	144
6.10	Linear Filters	146
6.11	Tying Up Loose Ends	150
6.12	Further Topics	152
6.13	Conceptual Questions	155
7	Introduction to Multivariate Methods	156
7.1	The Problem	157
7.2	Vectors	159
7.3	The Linear Transformation	160
7.4	Linear Independence	163
7.5	Matrix Operations	166
7.6	Invertible Transformations	168
7.7	Orthogonal Transformations	170
7.8	Random Vectors	172
7.9	Diagonalizing a Covariance Matrix	175
7.10	Multivariate Normal Distribution	178
7.11	Hotelling's T-squared Test	179
7.12	Multivariate Acceptance and Rejection Regions	181
7.13	Further Topics	182
7.14	Conceptual Questions	183
8	Linear Regression: Least Squares Estimation	185
8.1	The Problem	186
8.2	Method of Least Squares	188
8.3	Properties of the Least Squares Solution	192
8.4	Geometric Interpretation of Least Squares Solutions	196
8.5	Illustration Using Atmospheric CO ₂ Concentration	199
8.6	The Line Fit	205
8.7	Always Include the Intercept Term	206
8.8	Further Topics	207
8.9	Conceptual Questions	209
9	Linear Regression: Inference	210
9.1	The Problem	211
9.2	The Model	212
9.3	Distribution of the Residuals	212
9.4	Distribution of the Least Squares Estimates	213
9.5	Inferences about Individual Regression Parameters	215

9.6	Controlling for the Influence of Other Variables	216
9.7	Equivalence to “Regressing Out” Predictors	218
9.8	Seasonality as a Confounding Variable	222
9.9	Equivalence between the Correlation Test and Slope Test	224
9.10	Generalized Least Squares	225
9.11	Detection and Attribution of Climate Change	226
9.12	The General Linear Hypothesis	233
9.13	Tying Up Loose Ends	234
9.14	Conceptual Questions	236
10	Model Selection	237
10.1	The Problem	238
10.2	Bias–Variance Trade off	240
10.3	Out-of-Sample Errors	243
10.4	Model Selection Criteria	245
10.5	Pitfalls	249
10.6	Further Topics	253
10.7	Conceptual Questions	254
11	Screening: A Pitfall in Statistics	255
11.1	The Problem	256
11.2	Screening <i>iid</i> Test Statistics	259
11.3	The Bonferroni Procedure	262
11.4	Screening Based on Correlation Maps	262
11.5	Can You Trust Relations Inferred from Correlation Maps?	265
11.6	Screening Based on Change Points	265
11.7	Screening with a Validation Sample	268
11.8	The Screening Game: Can You Find the Statistical Flaw?	268
11.9	Screening Always Exists in Some Form	271
11.10	Conceptual Questions	272
12	Principal Component Analysis	273
12.1	The Problem	274
12.2	Examples	276
12.3	Solution by Singular Value Decomposition	283
12.4	Relation between PCA and the Population	285
12.5	Special Considerations for Climate Data	289
12.6	Further Topics	295
12.7	Conceptual Questions	297

Contents

ix

13	Field Significance	298
	13.1 The Problem	299
	13.2 The Livezey–Chen Field Significance Test	303
	13.3 Field Significance Test Based on Linear Regression	305
	13.4 False Discovery Rate	310
	13.5 Why Different Tests for Field Significance?	311
	13.6 Further Topics	312
	13.7 Conceptual Questions	312
14	Multivariate Linear Regression	314
	14.1 The Problem	315
	14.2 Review of Univariate Regression	317
	14.3 Estimating Multivariate Regression Models	320
	14.4 Hypothesis Testing in Multivariate Regression	323
	14.5 Selecting X	324
	14.6 Selecting Both X and Y	328
	14.7 Some Details about Regression with Principal Components	331
	14.8 Regression Maps and Projecting Data	332
	14.9 Conceptual Questions	333
15	Canonical Correlation Analysis	335
	15.1 The Problem	336
	15.2 Summary and Illustration of Canonical Correlation Analysis	337
	15.3 Population Canonical Correlation Analysis	343
	15.4 Relation between CCA and Linear Regression	347
	15.5 Invariance to Affine Transformation	349
	15.6 Solving CCA Using the Singular Value Decomposition	350
	15.7 Model Selection	357
	15.8 Hypothesis Testing	359
	15.9 Proof of the Maximization Properties	362
	15.10 Further Topics	364
	15.11 Conceptual Questions	364
16	Covariance Discriminant Analysis	366
	16.1 The Problem	367
	16.2 Illustration: Most Detectable Climate Change Signals	370
	16.3 Hypothesis Testing	378
	16.4 The Solution	382
	16.5 Solution in a Reduced-Dimensional Subspace	388
	16.6 Variable Selection	392

x	<i>Contents</i>	
	16.7 Further Topics	395
	16.8 Conceptual Questions	398
17	Analysis of Variance and Predictability	399
	17.1 The Problem	400
	17.2 Framing the Problem	401
	17.3 Test Equality of Variance	403
	17.4 Test Equality of Means: ANOVA	404
	17.5 Comments about ANOVA	406
	17.6 Weather Predictability	407
	17.7 Measures of Predictability	411
	17.8 What Is the Difference between Predictability and Skill?	414
	17.9 Chaos and Predictability	416
	17.10 Conceptual Questions	417
18	Predictable Component Analysis	418
	18.1 The Problem	419
	18.2 Illustration of Predictable Component Analysis	422
	18.3 Multivariate Analysis of Variance	424
	18.4 Predictable Component Analysis	427
	18.5 Variable Selection in PrCA	430
	18.6 PrCA Based on Other Measures of Predictability	432
	18.7 Skill Component Analysis	435
	18.8 Connection to Multivariate Linear Regression and CCA	437
	18.9 Further Properties of PrCA	439
	18.10 Conceptual Questions	445
19	Extreme Value Theory	446
	19.1 The Problem and a Summary of the Solution	447
	19.2 Distribution of the Maximal Value	453
	19.3 Maximum Likelihood Estimation	459
	19.4 Nonstationarity: Changing Characteristics of Extremes	463
	19.5 Further Topics	466
	19.6 Conceptual Questions	467
20	Data Assimilation	468
	20.1 The Problem	469
	20.2 A Univariate Example	469
	20.3 Some Important Properties and Interpretations	473
	20.4 Multivariate Gaussian Data Assimilation	475

<i>Contents</i>		xi
20.5	Sequential Processing of Observations	477
20.6	Multivariate Example	478
20.7	Further Topics	481
20.8	Conceptual Questions	487
21	Ensemble Square Root Filters	489
21.1	The Problem	490
21.2	Filter Divergence	497
21.3	Monitoring the Innovations	499
21.4	Multiplicative Inflation	500
21.5	Covariance Localization	503
21.6	Further Topics	507
21.7	Conceptual Questions	509
<i>Appendix</i>		510
A.1	Useful Mathematical Relations	510
A.2	Generalized Eigenvalue Problems	511
A.3	Derivatives of Quadratic Forms and Traces	512
<i>References</i>		514
<i>Index</i>		523

Cambridge University Press
978-1-108-47241-8 — Statistical Methods for Climate Scientists
Timothy DelSole , Michael Tippett
Frontmatter
[More Information](#)

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.

Cambridge University Press
978-1-108-47241-8 — Statistical Methods for Climate Scientists
Timothy DelSole , Michael Tippett
Frontmatter
[More Information](#)
