Wiley Series in Probability and Statistics

Statistics for SPATIO-TEMPORAL DATA

Noel Cressie • Christopher K. Wikle

WILEY

ftp:// SITE AVAILABLE

Preface

> *Nothing puzzles me more than time and space; and yet nothing troubles me less.…*

These words, by the English essayist Charles Lamb in his 1810 letter to Thomas Manning, provide a concise summary of this book. Of course, he was not thinking of Statistics, nor Science, nor Statistical Science, rather the more ephemeral notion of space and time. But here, in the physical world, there are scientific questions to resolve and predictions to make, and understanding the effects of dependencies across time and space is a crucial part. Those dependencies are in there somewhere, and we too are puzzled by them.

Up until the mid-twentieth century, Statistics' response to this puzzle was often to ignore the complicated structure, or to find clever ways to remove it. Interestingly, Charles Lamb seemed to be in tune with this, as he finished his sentence by saying, "as I never think about them." As Statistics has progressed through the twentieth century and moved into the twenty-first, spatial and temporal statistical methodology has been incorporated more and more into the scientific models of our world and indeed of our universe. As technology has improved, Statistics has been given a Rosetta Stone to begin to unlock Science's mysteries, from the molecular to the global to the cosmological.

In the beginning, there were data. Then, there were theories, formed from data, and those theories rose and fell according their agreement with new data. Data are central, and they should be cared for accordingly. To do good Science, databases should be fully documented and

algorithms (“black boxes”) involved in creating them should be in the public domain.

However, data alone do not tell us all that much about our world. When we look at data, how do we know if what we are seeing is “signal” as opposed to “noise”? How do we compare two sources of data—what is the basis of a comparison?

Similarly, theories (or models) by themselves are not often the best descriptions of the real world. What can be said about processes on scales at which there were no observations? What about uncertainties in parameters, or forcings, or interactions with other processes? Indeed, the key is the proper blending of such models and the data. Sometimes this might be done informally, for example, by taking a simulated field, simulated from a mathematical (say) model, and visually comparing it to a field of actual data. From the visual comparison, a deficiency in the model might be obvious, which might lead to a parameter adjustment, or even a new parameter. Then a *new* simulation might be implemented and a new visual comparison made, and so forth. This is one way to combine data and model, but it is not a very efficient way to deal with either. Indeed, the power of Statistical Science is that it provides several frameworks in which to combine data and model, in optimal ways, for the purpose of scientific inference. It might seem strange that there is not just one framework in which to carry out inferences, but even Statistics has its tribes. However, the element that is common to all is an attempt to partition variability and to quantify uncertainty.

In this book, we take the firm stand that the best paradigm (to date) in which to partition variability and quantify uncertainty is the *hierarchical statistical model.* Such a model explicitly acknowledges uncertainty in the data,

different from that in the process and parameters, and then it accommodates the uncertainty in the process (and finally in parameters, if necessary). We have used color in key places in our exposition, to distinguish between the parts of the hierarchical model concerned with data (green), with processes (blue) and with parameters (purple).

Hierarchical thinking (i.e., hierarchical modeling) is intimately tied to conditional thinking (i.e., modeling with conditional probabilities). Indeed, it is our perspective that conditional thinking is the aforementioned Rosetta Stone: It allows us to separately partition the effects of measurement error and scales of variability below the resolution of our data, conditioned on the process at possibly some other scale. Similarly, conditional thinking allows us to model a spatio-temporal process as it actually *evolves* through time, as opposed to just accounting for its marginal dependencies. Equally important, it allows us to use spatio-temporal dependencies in errors and/or parameters as a proxy for unknown and unknowable processes. Its impact goes even deeper, in that it allows parameters themselves to be dependent on other processes or other sources of data. Finally, it makes clear that if some components of variability are not interesting for a particular question, then the end result should *not* look like the data. Any unwanted components (e.g., measurement error) are filtered out: What you see (data) is not always what you want to get (process).

Critically, in the presence of data, conditional thinking allows sensible trade-offs to be made between data availability, process complexity, and computational complexity. Combined with the ever-expanding computational tools of the twenty-first century, the hierarchical statistical framework provides the structure to tackle important questions in Energy, Climate, Environment, Food, Finance, and so on. This will be a

century of massive (spatio-temporal) datasets collected to answer Society's dominant questions. In this book, we are particularly interested in inference for Climate and the Environment, where processes at small spatio-temporal scales influence those at larger scales, and vice versa. The questions to be resolved are fundamental to sustaining our planet, they involve complex spatio-temporal phenomena, *and* they are inherently statistical.

Our lives are spent marching through a space-time continuum, but space is different from time. We can (and often do) visit the same place over and over but always at different, ordered time points. We can go north, south, east, and west, up and down, but only ever forward from the past to the present and into the future. Any study of a spatio-temporal phenomenon needs to respect this difference. In this book, we have proposed spatio-temporal statistical methodologies that align with the underlying science, and we have found that hierarchical thinking is a natural way to achieve this alignment.

In the pages that follow, we have deliberately tried to build a bridge between the twentieth and twenty-first centuries in our presentation of spatio-temporal statistics. There are strong and powerful traditions that have developed in the last few decades and, even if they are not hierarchical, they provide the tools and motivation that can be used in hierarchical thinking. In some cases, hierarchical approaches that may be appealing in principle may be out of our reach in terms of timely implementation.

The official ftp site associated with the book can be found at:

ftp://ftp.wiley.com/public/sci_tech_med/spatio_temporal_data

In addition to posting errata, we will post supplemental material that we hope will be helpful to the reader.

As mentioned above, we have questions to resolve and predictions to make, so let’s get started….

NOEL CRESSIE
CHRISTOPHER K. WIKLE

Columbus, Ohio
Columbia, Missouri
December 2010

Acknowledgments

There are many people to thank, but two are missing. I wish my parents were alive to read this and know how important they are to what I do. I finally have a good answer to Ray's question, "How's the book going?"

My children, Amie and Sean, were only just present for my last book project, and now, as young adults, they are happily cheering on the completion of this one. Everyone has a book in them, and I hope they write theirs.

My dear friends and sibs may not understand all that is on these pages, but they know how important it was for me to "write it down." I would like them to know how important they are to me.

My co-author Chris has been a generous colleague and friend, and I admire enormously his intelligence and erudition. Together, we have been able to write a book that is so much more than what was in each of us; I have learned much from him. The order of authors is simply determined by the alphabet.

I have been taught by co-authors, colleagues, and students, who have shaped the material in this book. Their contributions are there in many, many ways. Some deserve special mention, which can be found in the joint acknowledgment.

The Department of Statistics at The Ohio State University has provided a nurturing environment for the whole lifetime of this book project. As part of my role as Director of the Department's Program in Spatial Statistics and Environmental Statistics, I have relied on my Program Assistant, Terry England, in many ways. This book has been a major project we have shared, and her efforts (which are

described in the joint acknowledgment) are deeply appreciated and warmly acknowledged. Paul Brower, our Department Administrative Manager, has been a great supporter of this book project and, more generally, of my attempts to “save the planet.” The office staff, the computer-support personnel, and the custodial staff are all warmly thanked for making my working environment in Cockins Hall a place where things get done … with good cheer.

Finally, it is not lost on me that this book will appear exactly 20 years after an earlier book I authored on spatial data. A lot can happen in two decades….

N. C.

My rather unorthodox training in Statistics and Meteorology would not have been possible without the tremendous support of my advisors, Tsing-Chang (Mike) Chen and Noel Cressie, both of whom set a great example of having mastered the art of teaching and the more elusive skill of mentoring. Mike Chen opened my eyes to the beauty of dynamics through his courses in Dynamic Meteorology and Geophysical Fluid Dynamics (in particular) and his patient one-on-one expositions; he is a master of explaining the “real” meaning behind the mathematics. In addition to teaching these subjects, he taught me the *philosophy of science* and to appreciate the importance of knowing the history of your field. Noel has been an inspirational mentor, collaborator, and friend. In addition to opening my eyes to the beauty of spatial statistics, he has, more than anyone, taught me to think like a statistician, and to stand up for the things in which I believe scientifically. At the time I was choosing an advisor in Statistics, some students warned me that I did not want to work with Noel because he was too “hard.” One of the best decisions I ever made professionally was to ignore that

advice! If hard means having high expectations, and providing generous advice and support, than I guess he was. Such a mentor has turned into a good friend. Noel mentioned at the beginning of this project how it can be very trying on a friendship to write a book. To be sure, we've had some disagreements, but we have managed quite well and I would not trade this journey for any other professional endeavor. I have learned a tremendous amount from Noel in writing this book, and I believe our friendship is stronger today than it was when we started.

I also want to thank two wonderful mentors with whom I started working when I was a postdoc at the National Center for Atmospheric Research (NCAR), and with whom I have continued to work along the way: Ralph Milliff and Mark Berliner. Ralph is a friend, a mentor, and an inspiration. He is a true scientist and is one of the most generous individuals I have met professionally. On every project in which we work, Ralph reminds me by the example he sets of what it means to be a scientist with integrity. Ralph saw the potential of Bayesian hierarchical modeling (BHM) as a way to manage uncertainty in oceanography before anyone else and, more importantly, he has stuck with it through good times and bad. Although he probably wouldn't admit it, Ralph is indeed an expert in BHM. Mark Berliner probably deserves a special chapter in this book. In fact, the book owes its philosophy to the ideas about hierarchical modeling and science that Mark began promoting when he was the director of the Geophysical Statistics Project (GSP) at NCAR in the mid-1990s. My time in that group was the most stimulating intellectual environment of my career. Mark had a vision of how BHM could be used as a paradigm for Science. Indeed, Mark is a master of weaving Science into the fabric of Statistics; he is always an inspiration! Perhaps more importantly, Mark is a friend and mentor and is one of the truly generous people

in Statistics. There is no doubt that I would not be the statistician I am today if it were not for Mark's mentoring and collaborations.

A special thanks to Andy Royle, who was also part of the GSP group. Andy is an amazing statistician, and our interactions and collaborations during my time at NCAR and immediately thereafter were very important in shaping the way in which I envisioned Mark's philosophy playing a role in spatio-temporal statistics. In fact, Andy and I talked about writing a lower-level book on spatio-temporal statistics while we were at NCAR, and we even drafted an outline and a couple of chapters. We both moved in different directions (Andy to ecological statistics in the Fish and Wildlife Service and then to the U.S. Geological Survey, and me to academia), which prevented us from pursuing that project. But, there is no doubt that there is much of Andy in this book.

My path to Statistics was certainly not direct. In fact, the first classes I had in the subject as an undergraduate did little to impress upon me its importance. However, along the way I have had the great fortune of having some fantastic teachers and collaborators. In particular, I would like to thank Peter Sherman and Rol Madden, for introducing me to the power of spectral analysis and its scientific interpretation; and Mark Kaiser, for introducing me to hierarchical modeling in the early 1990s through a wonderful experimental course at Iowa State University. Many other people have influenced me along the way, either through conversations or collaborations. In particular, knowing that I am sure to unintentionally leave someone off of this list, I would like to thank: Chris Anderson, Thomas Bengtsson, Jim Clark, Bob Dorazio, William Dunsmuir, Dave Higdon, Tim Hoar, Dave Larsen, Andy Moore, Doug Nychka, Nadia Pinardi, Yanyan Sheng, Jon Stroud, Joe Tribbia, and Jay Ver Hoef. Of course, there

are many others who have influenced me with the quality of their research, their integrity, and through brief discussions at meetings.

I would also like to thank my friends, colleagues, and students at the University of Missouri, past and present. They have been there for me no matter what, and they have made it a pleasure to come to the office. I would like to thank Joe Cavanaugh, Wade Davis, Neil Fox, Scott Holan, Sakis Micheas, Jake Oleson, Larry Ries, Thomas Rose, and Mark Wildhaber, for their much-valued friendship. In particular, the long-standing friendship of Larry Ries and Joe Cavanaugh has been a source of strength and an invaluable outlet; thanks for listening, guys! As a new Assistant Professor, I was so fortunate to have a mentor like Joe, who sets the “gold standard” of what it means to be an academic. Furthermore, I have been fortunate to have a wonderful group of Master’s and Ph.D. students over the years, all of whom have contributed to my view of Statistics. My Ph.D. students, past and present, have kept me on my toes, and I value their collaboration and their friendship. Much of what is in this book comes from what we did together. In particular, I would like to acknowledge the implicit contributions of Ali Arab, Mevin Hooten, Yong Song, and Ke (Bill) Xu.

My biggest thanks go to my family! I am eternally grateful to my parents, Bayliss and Irene Wikle, for instilling in me the value of education and for giving me the freedom and support to pursue my dreams. Their support and love has shown me what it means to be a parent. In addition, I want to thank my brothers Shawn, Jeff, and Tim for everything they have done for me through the years. There is no doubt that they got all the brains in the family! I am sure that I continue to benefit in my professional life from the friendly competitions and creative activities that we engaged in, growing up on the “Wikle ancestral compound.” In

addition, I would like to thank my in-laws, George and MaryIda Heskamp, for their support through the years and for further strengthening my sense of family. Most importantly, I want to thank Carolyn, Olivia, Nathan, and Andrea. Your patience and support while I was working on this project have been a source of strength, and I deeply appreciate it. Olivia, Nathan, and Andrea, you have taught me much more than any book or research paper ever could, and you have provided countless sources of inspiration. Carolyn, you have supported my career unconditionally, and I am so fortunate to be able to share my life and love with you. There is no doubt that I am a better person every single day because of you! Thank you.

C. K. W.

There are many people who deserve our thanks for their direct contributions to this book. At Ohio State and beyond, Rajib Paul has generously prepared many of our figures, well beyond his time as a student; Matthias Katzfuß helped in a pivotal way with his timely reading and editing of most of the chapters; Bethann Pflugeisen was our “reference and editing maven,” as well as helping with the rhyme and reason in Chapters 1 and 2; Ana Lucía Ortiz of Universidad Francisco Marroquín helped us obtain permission to use the lienzo shown on the front cover and in Chapter 1; Gardar Johannesson and Desheng Liu made figures in Chapter 4; and Emily Kang made figures in Section 9.2. At Missouri and beyond, Scott Holan gave comments on an early draft of Chapter 3; Sakis Micheas gave comments on Section 4.4.3; Mevin Hooten made figures and gave comments on Section 9.3; Ralph Milliff made figures and gave comments on Section 9.4; Jeremiah Brown, Nadia Pinardi, Alessandro Bonazzi, and the INGV group contributed to figures in Section 9.4; Ali Arab, Rima Dey, Dan Gladish, Mevin Hooten, Bill Leeds, and Wen-Hsi Yang gave general comments on drafts of various chapters; and

Mary Peng, Dan Gladish, and Carolyn Wikle helped with the index.

Finally, we wish to give special thanks to Terry England for her invaluable help in typing large portions of the manuscript and for her tremendous LaTeX skills in organizing and typesetting the book. We could not have finished this without her!

At Wiley, the acquisition, production (particularly Lisa Van Horn), and marketing teams have all played a crucial role in taking our idea to write a book on spatio-temporal statistics, all the way to a four-color volume in the Wiley Series in Probability and Statistics. This is a publisher that is committed to getting our ideas into the offices, labs, and libraries of scientists and engineers. We want to change the way they practice Statistics … particularly for spatio-temporal data.

N. C.
C. K. W.

CHAPTER 1
Space–Time: The Next Frontier

This book is about the statistical analysis of data … *spatio-temporal* data. By this we mean data to which labels have been added showing where and when they were collected. Good science protocol calls for data records to include place and time of collection. Causation is the “holy grail” of Science, and hence to infer cause–effect relationships (i.e., “why”) it is essential to keep track of “when”; a cause always precedes an effect. Keeping track of “where” recognizes the importance of knowing the “lay of the land”; and, quite simply, there would be no History without Geography.

We believe that in order to answer the “why” question, Science should address the “where” and “when” questions. To do that, spatio-temporal datasets are needed. However, spatial datasets that do not have a temporal dimension can occur in many areas of Science, from Archeology to Zoology. The spatial data may be from a “snapshot” in time (e.g., liver-cancer rates in U.S. counties in 2009), or they may be taken from a process that is not evolving in time (e.g., an iron-ore body in the Pilbara region of Australia). Sometimes, the temporal component has simply been discarded, and the same may have happened to the spatial component as well. Also, temporal datasets that do not have a spatial dimension are not unusual, for analogous reasons. For example, two time series, one of monthly mean carbon dioxide measurements from the Mauna Loa Observatory, Hawaii, and the other of monthly surface temperatures averaged across the globe, do not have a spatial dimension (for different reasons).

Spatio-Temporal Data

Spatio-temporal data were essential to the nomadic tribes of early civilization, who used them to return to seasonal hunting grounds. On a grander scale, datasets on location, weather, geology, plants, animals, and indigenous people were collected by early explorers seeking to map new lands and enrich their kings and queens. The conquistadors of Mesoamerica certainly did this for Spain.

The indigenous people also made their own maps of the Spanish conquest, in the form of a *lienzo.* A lienzo represents a type of historical cartography, a painting on panels of cloth that uses stylized symbols to tell the *history* of a *geographical* region. The *Lienzo de Quauhquechollan* is made up of 15 joined pieces of cotton cloth and is a map that tells the story, from 1527 to 1530, of the Spanish conquest of the region now known as Guatemala. It has been restored digitally in a major project by Exploraciones sobre la Historia at the Universidad Francisco Marroquín (UFM) in Guatemala City (see Figure 1.1). This story of the Spanish conquest in Guatemala is an illustration of complex spatio-temporal interactions. Reading the lienzo and understanding its correspondence with the geography of the region required deciphering; see Asselbergs (2008) for a complete description. The original lienzo dates from about 1530 and represents a spatio-temporal dataset that is almost 500 years old!

In a sense, we are all analyzers of spatial and temporal data. As we plan our futures (economically, socially, academically, etc.), we must take into account the present and seek guidance from the past. As we look at a map to plan a trip, we are letting its spatial abstraction guide us to our destination. The philosopher Ludwig Wittgenstein compared language to a city that has evolved over time (Wittgenstein, 1958): “Our language can be seen as an

ancient city: A maze of little streets and squares, of old and new houses, and of houses with additions from various periods; and this surrounded by a multitude of new burroughs with straight and regular streets and uniform houses!”

Graphs of data indexed by time (time series) and remote-sensing images made up of radiances indexed by pixel location (spatial data) show variability at a glance. For example, Figure 1.2 shows the Missouri River *gage-height* levels during the 10-year period, 1988–1997, at Hermann, MO. Figure 1.3 shows two remotely sensed images of the river taken in September 1992, before a major flood event, and in September 1993, after the highest crest ever recorded at Hermann (36.97 ft on July 31, 1993). The top panel of Figure 1.3 shows the town of Gasconade in the middle of the scene, situated in the “V” where the Gasconade River joins the Missouri River; Gasconade is at mile 104.4 and eight miles downstream is the river town of Hermann, visible at the very bottom of the scenes. Notice the intensive agriculture in the river’s flood plain in September 1992. The bottom panel of Figure 1.3 shows the same region, one year later, after the severe flooding in the summer of 1993. The inundation of Gasconade, the floodplain, and the environs of Hermann is stunning. There is a multiscale process behind all of this that involves where, when, and how much precipitation occurred upstream, the morphology of the watershed, microphysical soil properties that determine run-off, the U.S. Army Corps of Engineers’ construction of levees upstream, and so on. However, by looking only in the spatial dimension, or only in the temporal dimension, we miss the dynamical evolution of the flood event as it progressed downstream. Spatio-temporal data on this portion of the Missouri River, which shows how the river got from “before” to “after,” would be

best illustrated with a movie, showing a temporal sequence of spatial images before, during, and after the flood.

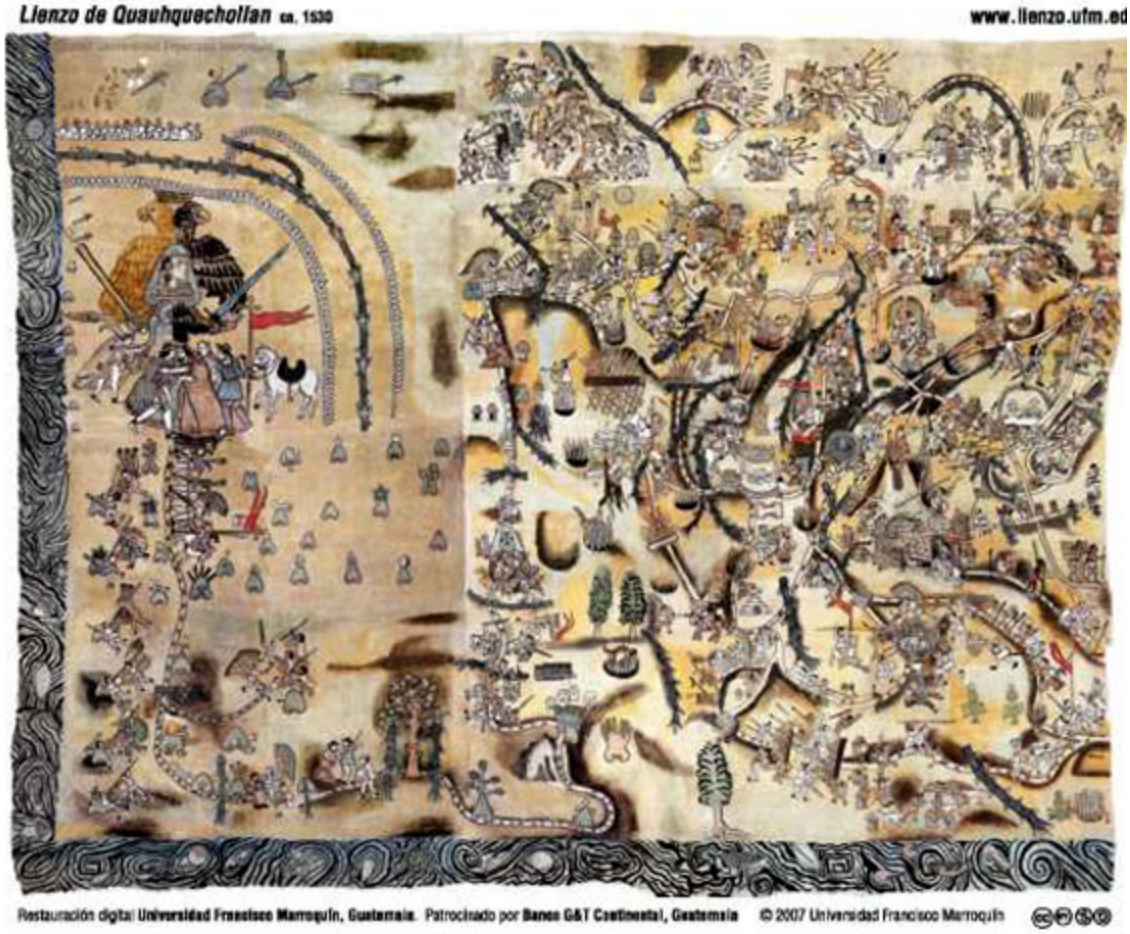

Figure 1.1 Digitally restored Lienzo de Quauhquechollan, whose actual dimensions are 2.45 m in height by 3.20 m in width. [Image is available under the Creative Commons license Attribution-Noncommercial-Share Alike © 2007 Universidad Francisco Marroquín.]

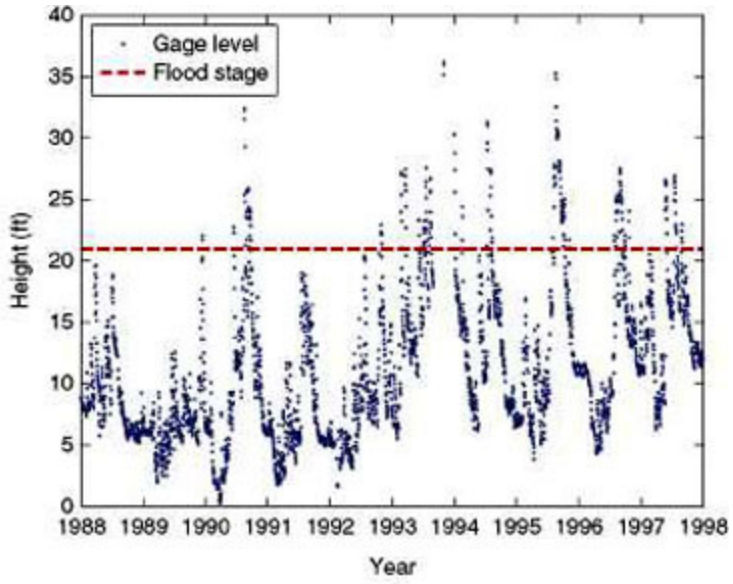

Figure 1.2 Time-series levels of gage height at Hermann, MO (mile 96.5 on the Missouri River) from January 1, 1988 through December 31, 1997. Flood stage is given by the horizontal dashed line. The highest recorded gage height in the 10-year period was 36.97 ft on July 31, 1993.

There is an important statistical characteristic of spatio-temporal data that is very common, namely that nearby (in space and time) observations tend to be more alike than those far apart. However, in the case of "competition," the opposite may happen (e.g., under big trees only small trees can grow), but the general conclusion is nevertheless that spatio-temporal data should *not* be modeled as being statistically independent. [Tobler (1970) called this notion "the first law of Geography."] Even if spatio-temporal trends are used to capture the dependence at large scales, there is typically a cascade of smaller spatio-temporal scales for which a statistical model is needed to capture the dependence. Consequently, an assumption that spatio-temporal data follow the "independent and identically distributed" *(iid)* statistical paradigm should typically be avoided. Paradigms that incorporate dependence are needed: The time series models in Chapter 3 and the

spatial process models in Chapter 4 give those paradigms for *temporal data* and *spatial data,* respectively. From Chapter 5 onwards, we are concerned directly with Statistics for *spatio-temporal data.*

Uncertainty and the Role of Statistics

Uncertainty is everywhere; as Benjamin Franklin famously said (Sparks, 1840), “In this world nothing can be said to be certain, except death and taxes.” Not only is our world uncertain, our attempts to explain the world (i.e., Science) are uncertain. And our measurements of our (uncertain) world are uncertain. Statistics is the “Science of Uncertainty,” and it offers a coherent approach to handling the sources of uncertainty referred to above. Indeed, in our work we use the term *Statistical Science* interchangeably with *Statistics* (with a “capital” S); we use *statistics* (with a “small” s) to refer to summaries of the data.

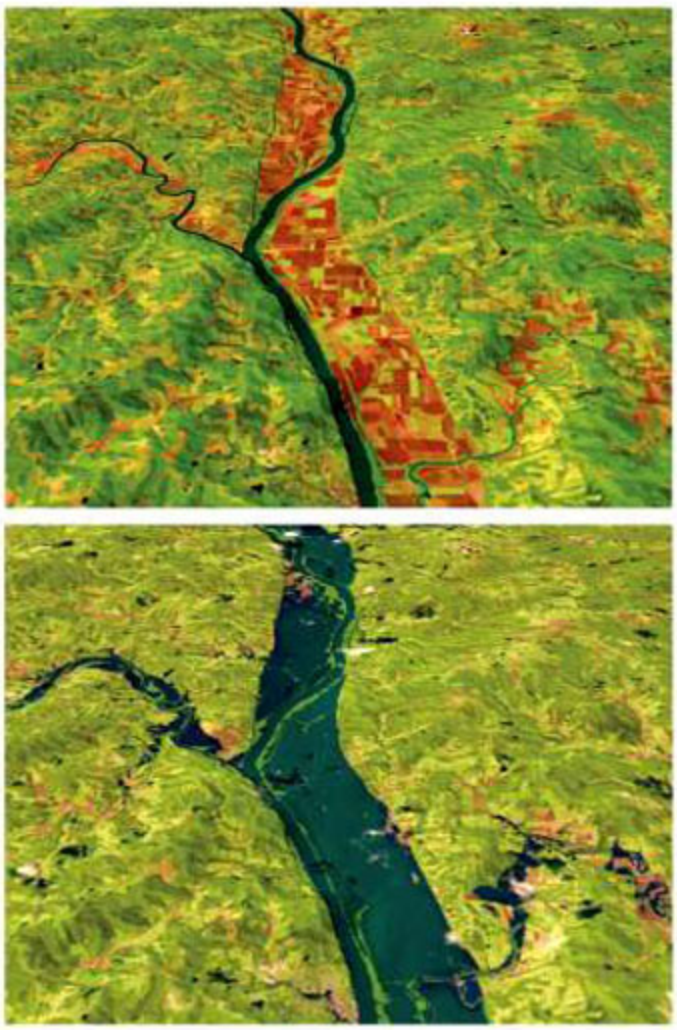

Figure 1.3 Images from NASA's Landsat Thematic Mapper. Each image shows a segment of the Missouri River near Hermann, MO (mile 96.5, at the bottom of the scene), and Gasconade, MO (mile 104.4, in the "V" in the middle of the scene). The river flows from west (top of the scene) to east (bottom of the scene). **Top panel:** September 1992, before a major flood event. **Bottom panel:** September 1993, after a record-breaking flood event in July 1993.

In most of this book, we shall express uncertainty through variability, but we note that other measures (e.g., entropy) could also be used. Just as the physical and biological sciences have the notions of mass balance and energy

balance, Statistical Science has a notion of variability balance. The total variability is modeled with variability due to *measurement,* variability due to using a (more-or-less uncertain) *model* of how the world works, and variability due to uncertainty on *parameters* that control the measurement and model variabilities.

Although real-world systems may in principle be partially deterministic, our information is incomplete at each of the stages of observation, summarization, and inference, and thus our understanding is clouded by uncertainty. Consequently, by the time the inference stage is reached, the lack of certainty will influence how much knowledge we can gain from the data. Furthermore, if the dynamics of the system are nonlinear, the processes can exhibit *chaos* (Section 3.2.4), even though the theory is based on *deterministic* dynamical systems. (In Chapters 3 and 7, we show how model uncertainty in these systems naturally leads to *stochastic* dynamical systems that incorporate *system,* or *intrinsic,* noise.)

Data can hold so much potential, but they are an entropic collection of digits or bits unless they can be organized into a database. With the ability in a database to structure, search, filter, query, visualize, and summarize, the data begin to contain *information.* Some of this information comes from judicious use of statistics (i.e., summaries) with a "small s." Then, in going from information to *knowledge,* Science (and, with it, Statistics with a "capital S") takes over. This book makes contributions at all levels of the data–information–knowledge pyramid, but we generally stop short of the summit where knowledge is used to determine policy. The methodology we develop is poised to do so, and we believe that at the interface between Science, Statistics, and Policy there is an enormous need for (spatio-temporal) decision-making in the presence of uncertainty.

In this book, we approach the problem of “scientific understanding in the presence of uncertainty” from a probabilistic viewpoint, which allows us to build useful spatio-temporal statistical models and make scientific inferences for various spatial and temporal scales. Accounting for the uncertainty enables us to look for possible associations within and between variables in the system, with the potential for finding mechanisms that extend, modify, or even disprove a scientific theory.

Uncertainty and Data

Central to the observation, summarization, and inference (including prediction) of spatio-temporal processes are *data.* All data come bundled with error. In particular, along with the obvious errors associated with measuring, manipulating, and archiving, there are other errors, such as discrete spatial and temporal sampling of an inherently continuous system. Consequently, there are always scales of variability that are unresolvable and that will further “contaminate” the observations. For example, in Atmospheric Science, this is considered a form of “turbulence,” and it corresponds to the well known aliasing problem in time series analysis (e.g., see Section 3.5.1; Chatfield, 1989, p. 126) and the microscale component of the “nugget effect” in geostatistics [e.g., see the introductory remarks to Chapter 4 and Cressie (1993, p. 59)].

Furthermore, spatio-temporal data are rarely sampled at spatial or temporal locations that are optimal for the analysis of a specific scientific problem. For instance, in environmental studies there is often a bias in data coverage toward areas where population density is large, and within a given area the coverage may be limited by cost. Thus, the location of a measuring site and its temporal sampling frequency may have very little to do with the underlying

scientific mechanisms. A scientific study should include the *design* of data locations and sampling frequencies when framing questions, when choosing statistical-analysis techniques, and when interpreting results. This task is complicated, since the data are nearly always statistically dependent in space and time, and hence most of the traditional statistical methods taught in introductory statistics courses (which assume *iid* errors) do not apply or have to be modified.

Uncertainty and Models

Science attempts to *explain* the world in which we live, but that world is very complex. A model is a simplification of some well chosen aspects of the world, where the level of complexity often depends on the question being asked. Pragmatically, the goal of a model is to predict, and at the same time scientists want to incorporate their understanding of how the world works into their models. For example, the motion of a pendulum can be modeled using Newton's second law and the simple gravity pendulum that ignores the effect of friction and air resistance. The model predicts future locations of the pendulum quite well, with smaller-order modifications needed when the pendulum is used for precise time-keeping. Models that are scientifically meaningful, that predict well, and that are conceptually simple are generally preferred. An injudicious application of Occam's razor (or "the law of parsimony") might elevate simplicity over the other two criteria. For example, a statistical model based on correlational associations might be simpler than a model based on scientific theory. The way to bridge this divide is to focus on what is more or less certain in the scientific theory and use *scientific-statistical* relationships to characterize it.

Albert Einstein said: “It can scarcely be denied that the supreme goal of all theory is to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience,” at the Herbert Spencer Lecture delivered at Oxford University on June 10, 1933; see Einstein (1934). Much later, in the October 1977 issue of the *Reader’s Digest,* it appears as if Einstein’s quote was paraphrased to: “Everything should be made as simple as possible, but not simpler.” Statistics and its models, including those involving scientific–statistical relationships, should not be spared from following this advice. Royle and Dorazio (2008, pp. 414–415) give a succinct discussion of this desire for conceptual simplicity in a model. As the data become more expansive, it is natural that they might suggest a more complex model. Clearly, there is a balance to be struck between too much simplicity, and hence failing to recognize an important signal in the data, and too much complexity, which results in a nonexistent signal being “discovered.” One might call this desire for balance the *Goldilocks Principle* of modeling. *(Goldilocks and the Three Bears* is a nursery tale about a little girl’s discovery of what is “just right.”)

It is our belief that statistical models used for describing temporal variability in space should represent the variability dynamically. Models used in Physics, Chemistry, Biology, Economics, and so on, do this all the time with difference equations and differential equations to express the evolutionary mechanisms. Why should this change when the models become statistical? Perhaps it is because there is often an alternative—for example, a model based on autocorrelations that describe the temporal dependence. However, this descriptive approach does not directly involve evolutionary mechanisms and, as a consequence, it pushes understanding of the

Physics/Chemistry/Biology/Economics/etc. into the background. As has been discussed above, there is a way to have both, in the form of a scientific–statistical model that recognizes the dynamical scientific aspects of the phenomenon, with their *uncertainties* expressed through *statistical* models. Descriptive (correlational) statistical models do have a role to play when little is known about the etiology of the phenomenon; this approach is presented in Sections 6.1 and 6.2. Thereafter, this book adopts a dynamical approach to Statistics for spatio-temporal data.

Nearby Things Tend to Be More Alike…

A simple and sometimes effective forecast of tomorrow's weather is to use today's observed weather. This "persistence" forecast is based on observing large autocorrelations between successive days. Such dependence behavior in "nearby" temporal data is also seen in "nearby" spatial data, such as in studies of the environment. Statistics for Spatio-Temporal Data presents the next frontier; this book steps forward into new territories and revisits old ones. It reviews and extends different aspects of statistical methodology based on spatio-temporal dependencies: exploratory data analysis, marginal/conditional models in discrete/continuous time, optimal inference (including parameter estimation and process prediction), model diagnostics and evaluation, and so forth. One fundamental scientific problem that arises is understanding the evolution of processes over time, particularly in environmental studies (e.g., the evolution of sea-ice coverage in the Arctic; changes in sea level; time trends in precipitation). Proper inference to determine if evolutionary components (natural or anthropogenic) are real requires a spatio-temporal *statistical* methodology.

The scientific method involves observation, inspiration, hypothesis generation, experimentation (to support or

refute the current scientific hypothesis), inference, more inspiration, more hypothesis generation, and so forth. In a sense, everything begins with observation, but it is quickly apparent to a scientist that unless data are obtained in a more-or-less controlled manner (i.e., using an experimental design), proper inference can be difficult. This is the fundamental difference between “observation” and “experimentation.” Understanding the role of dependencies when the data are spatial or temporal, or both, provides an important perspective on working with experimental data versus observational data.

Experimental Data

Earth’s population is many billions, and the demand for sustenance is great and continuous. The planet’s ability to produce food on a massive scale largely came from fundamental experiments in crop science in the early twentieth century. Fisher (1935) developed a statistical theory of experimental design, based on the three principles of blocking, randomization, and replication, for choosing high-yielding, insect-resistant crops adapted to local conditions. He developed a vocabulary that is used today in scientific experiments of all types: response (e.g., wheat yields), treatments (e.g., varieties of wheat), factors (e.g., soil type, field aspect, growing season), levels of factors (e.g., for the soil-type factor, the levels might be sand, gravel, silt, clay, peat), plot (experimental unit that receives a single treatment), block (collection of plots with the same factor/level combination), randomization (random assignment of treatments to plots), replication (number of responses per treatment), and so on.

Data from designed experiments, when analyzed appropriately, allow stronger (almost) causative inferences, which incubate further scientific inspiration and hypothesis generation, and so forth, through the cycle. In the right

hands, and with a component of luck, this cycle leads to great breakthroughs [e.g., the discovery of penicillin in 1928 by Alexander Fleming; see, e.g., Hare (1970)]. Even small breakthroughs are bricks that are laid on the knowledge pyramid.

Space and time are fundamental factors of any experiment. For example, “soil type” is highly spatial and “growing season” is highly temporal. Protocol for any well designed experiment should involve recording the location and time at which each datum was collected, because so many factors (known or unknown) correlate with them. After the experiment has been performed, spatial and temporal information can be used as proxies for unknown, unaccounted-for factors that may later become “known” as the experiment proceeds. From this point of view, the natural place to put spatial and temporal effects in the statistical model is in the mean. But, there is an alternative ….

In R. A. Fisher’s pathbreaking work on design of experiments in agricultural science, he wrote (Fisher, 1935, p. 66): “After choosing the area we usually have no guidance beyond the widely verified fact that patches in close proximity are commonly more alike, as judged by the yield of crops, than those which are further apart.” Spatial variability, which to Fisher came in the form of plot-to-plot variability, is largely due to physical properties of the soil and environmental properties of the field. Fisher avoided the confounding of treatment effect with plot effect with the inspirational introduction of *randomization* into the scientific method. It was a brilliant insertion of *more* uncertainty into a place in the experiment where uncertainty abounds, leaving the more certain parts of the experiment intact. Fisher’s idea has had an enormous effect on all our lives. For example, any medicine we have taken to treat our ailments and illnesses has gone through

rigorous testing, to which the *randomized* clinical trial is central (where a “plot” is often the patient).

Randomization comes with a price. It allows valid inference on the treatments through a simple expression for the mean response, but the variances and covariances of the responses are affected too. Under randomization of the assignment of treatments to plots, the notions of “close proximity” and “far apart” have been hustled out the back door. Can we get spatial dependence back into the statistical analysis of responses, resulting in more efficient inferences for treatment effects? The answer to this is a resounding “Yes”; see the introduction to Chapter 4.

Observational Data

Organisms are born, live, reproduce, and die, but they can produce harmful by-products that may threaten their own well-being as well as the well-being of other organisms around them. (The species *Homo sapiens* is unique in many of its abilities, including its ability to have a major impact on all other organisms on Earth.) Variability within organisms can be large (e.g., within *H. sapiens*), as can variability between their environments. Thus, it can be very difficult to conduct controlled experiments on Earth’s ecology and environment.

Observational data come from a “wilder side” of Science. The environment (such as climate, air and water quality, radioactive contamination, etc.) is a part of our lives that often will not submit to blocking, randomization, and replication. We cannot control when it rains, nor can we observe two Los Angeles, one with smog and one without. We *can* look for two like communities, one with contaminated water and one without; and we *can* look at health records before and after a toxic emission. However, any inference is tentative because the two factors, space

and time, are not controlled for. Collecting samples from ambient air presents a philosophical problem because the parcel of air is unique when it passes the monitoring site; it evolves as the changes in air pressure move it around, and it will never come back to allow us the luxury of obtaining an independent, identically distributed observation. (If these observations are used to study the effect of air quality on human health, there is the further problem that the ambient air is not actually what individuals breathe in their homes or their workplaces; this introduces even more uncertainty into the study.)

In the environmental and life sciences, classical experimental design can struggle to keep up with the questions being asked, but they still need to be answered. And, as we have discussed just above, *uncertainty* is likely to be higher without experimental control. Thus, Statistical Science has a crucial role to play, although it does not fit neatly into the blocking–randomization–replication framework. Even when one is able to "block" the human subjects on age and sex, say, it may be that an unknown genetic factor will determine how a patient responds to a given treatment. (Personalized medicine has as one of its goals to make the unknown genetic factor known.) In epidemiological studies, controls may be randomly matched with cases, but the cases are in no way assigned randomly to neighborhoods. And, although duplicate chemical assays allow for assessment of measurement error in a study on stream pollution, replication of a water parcel from the stream is impossible. In such circumstances, Statistics is even more relevant, and we advocate that the scientific method invoke the principle of *expressing uncertainty through probabilities.*

In the environmental sciences, proximity in space and time is a particularly relevant factor. The word "environ" means "around" in French. While ecology is the study of

organisms, the *environment* is the surroundings of organisms. “Nearby” is a relative notion, relative to the spatial and temporal scales of the phenomenon under study. For example, in the spatial case, a toxic-waste-disposal site may directly affect a neighborhood of a few square miles; a coal-burning power plant may directly affect a heavily populated region of many tens of square miles, and an increase in greenhouse gases will affect the whole planet. Clearly, a global effect is felt locally in many ways, from a longer growing season in Alberta, Canada, to a redistribution of beachfront property in Florida, USA. The point we wish to make here is that a quantity like global mean temperature is a largely uninformative summary of how daily lives of a community will be affected by a warmer planet, which means that environmental studies of the globe must recognize the importance of *local* variability. Furthermore, how the spatial variability behaves dynamically (i.e., the spatio-temporal variability) is key to understanding the causes of global warming and what to do about it. Finally, we state the obvious, that political boundaries cannot hold back a one-meter rise in sea level; our environment is ultimately a global resource and its stewardship is an international responsibility.

Einsteinian Physics

Einstein’s theory of relativity (e.g., Bergmann, 1976) demonstrated that space and time are interdependent and inseparable. In contrast, our book is almost exclusively concerned with phenomena that reside in a classical Newtonian framework (e.g., Giancoli, 1998). We include a brief discussion of space and time within Einstein’s framework, to indicate that modifications would be needed for, say, spatio-temporal astronomical data.

Einstein proposed a “thought experiment,” a version of which we now give. Think of a boxcar being pulled by a