Chapter 1

Introduction

1.1. WHAT IS STATISTICS?

"Statistics is the discipline concerned with the study of variability, with the study of uncertainty, and with the study of decision-making in the face of uncertainty" (Lindsay et al. 2004, p. 388). This book is concerned with the use of statistical methods in the atmospheric sciences, specifically in the various specialties within meteorology and climatology, although much of what is presented is applicable to other fields as well.

Students (and others) often resist statistics, and many perceive the subject to be boring beyond description. Before the advent of cheap and widely available computers, this negative view had some basis, at least with respect to applications of statistics involving the analysis of data. Performing hand calculations, even with the aid of a scientific pocket calculator, was indeed tedious, mind numbing, and time consuming. The capacity of an ordinary personal computer is now well beyond the fastest mainframe computers of just a few decades ago, but some people seem not to have noticed that the age of computational drudgery in statistics has long passed. In fact, some important and powerful statistical techniques were not even practical before the abundant availability of fast computing, and our repertoire of these "big data" methods continues to expand in parallel with ongoing increases in computing capacity. Even when liberated from hand calculations, statistics is sometimes still seen as uninteresting by people who do not appreciate its relevance to scientific problems. Hopefully, this book will help provide that appreciation, at least within the atmospheric sciences.

Fundamentally, statistics is concerned with uncertainty. Evaluating and quantifying uncertainty, as well as making inferences and forecasts in the face of uncertainty, are all parts of statistics. It should not be surprising, then, that statistics has many roles to play in the atmospheric sciences, since it is the uncertainty about atmospheric behavior that makes the atmosphere interesting. For example, many people are fascinated by weather forecasting, which remains interesting precisely because of the uncertainty that is intrinsic to the problem. If it were possible to make perfect forecasts or nearly perfect forecasts even one day into the future (i.e., if there were little or no uncertainty involved), the practice of meteorology would present few challenges, and would be similar in many ways to the calculation of tide tables.

1.2. DESCRIPTIVE AND INFERENTIAL STATISTICS

It is convenient, although somewhat arbitrary, to divide statistics into two broad areas: descriptive statistics and inferential statistics. Both are relevant to the atmospheric sciences.

The descriptive side of statistics pertains to the organization and summarization of data. The atmospheric sciences are awash with data. Worldwide, operational surface and upper-air observations are

Statistical Methods in the Atmospheric Sciences. https://doi.org/10.1016/B978-0-12-815823-4.00001-8

routinely taken at thousands of locations in support of weather forecasting activities. These are supplemented with aircraft, radar, profiler, and satellite data. Observations of the atmosphere specifically for research purposes are less widespread, but often involve very dense sampling in time and space. In addition, dynamical models of the atmosphere,[1] which undertake numerical integration of the equations describing the physics of atmospheric flow, produce yet more numerical output for both operational and research purposes.

As a consequence of these activities, we are often confronted with extremely large batches of numbers that, we hope, contain information about natural phenomena of interest. It can be a nontrivial task just to make some preliminary sense of such data sets. It is typically necessary to organize the raw data, and to choose and implement appropriate summary representations. When the individual data values are too numerous to be grasped individually, a summary that nevertheless portrays important aspects of their variations—a statistical model—can be invaluable in understanding them. It is worth emphasizing that it is not the purpose of descriptive data analyses to "play with numbers." Rather, these analyses are undertaken because it is known, suspected, or hoped that the data contain information about a natural phenomenon of interest, which can be exposed or better understood through the statistical analysis.

Inferential statistics is traditionally understood as consisting of methods and procedures used to draw conclusions regarding underlying processes that generate the data. For example, one can conceive of climate as the process that generates weather (Stephenson et al., 2012), so that one goal of climate science is to understand or infer characteristics of this generating process on the basis of the single sample realization of the atmospheric record that we have been able to observe. Thiébaux and Pedder (1987) express this point somewhat poetically when they state that statistics is "the art of persuading the world to yield information about itself." There is a kernel of truth here: Our physical understanding of atmospheric phenomena comes in part through statistical manipulation and analysis of data. In the context of the atmospheric sciences, it is sensible to interpret inferential statistics a bit more broadly as well and to include statistical forecasting of both weather and climate. By now this important field has a long tradition and is an integral part of operational forecasting at meteorological centers throughout the world.

1.3. UNCERTAINTY ABOUT THE ATMOSPHERE

The notion of uncertainty underlies both descriptive and inferential statistics. If atmospheric processes were constant, or strictly periodic, describing them mathematically would be easy. Weather forecasting would also be easy, and meteorology would be boring. Of course, the atmosphere exhibits variations and fluctuations that are irregular. This uncertainty is the driving force behind the collection and analysis of the large data sets referred to in the previous section. It also implies that weather forecasts are inescapably uncertain. The weather forecaster predicting a particular temperature on the following day is not at all surprised (and perhaps is even pleased) if the subsequent observation is different by a degree or two, and users of everyday forecasts also understand that forecasts involve uncertainty (e.g., Joslyn and Savelli, 2010). "Uncertainty is a fundamental characteristic of weather, seasonal climate, and

1. These are often referred to as NWP (numerical weather prediction) models, which term was coined in the middle of the last century (Charney and Eliassen, 1949) in order to distinguish dynamical from traditional subjective (e.g., Dunn, 1951) weather forecasting. However, as exemplified by the contents of this book, statistical methods and models are also numerical, so that the more specifically descriptive term "dynamical models" seems preferable.

hydrological prediction, and no forecast is complete without a description of its uncertainty" (National Research Council, 2006). Communicating this uncertainty promotes forecast user confidence, helps manage user expectations, and honestly reflects the state of the underlying science (Gill et al., 2008).

In order to deal quantitatively with uncertainty it is necessary to employ the tools of probability, which is the mathematical language of uncertainty. Before reviewing the basics of probability, it is worthwhile to examine why there is uncertainty about the atmosphere. After all, we have large, sophisticated dynamical computer models that represent the physics of the atmosphere, and such models are used routinely for forecasting its future evolution. Individually, these models have traditionally been formulated in a way that is deterministic, that is, without the ability to represent uncertainty. Once supplied with a particular initial atmospheric state (pressures, winds, temperatures, moisture content, etc., comprehensively through the depth of the atmosphere and around the planet) and boundary forcings (notably solar radiation, and sea- and land-surface conditions), each will produce a single particular result. Rerunning the model with the same inputs will not change that result.

In principle, dynamical atmospheric models could provide forecasts with no uncertainty, but they do not, for two reasons. First, even though the models can be very impressive and give quite good approximations to atmospheric behavior, they do not contain complete and true representations of the governing physics. An important and essentially unavoidable cause of this problem is that some relevant physical processes operate on scales too small and/or too fast to be represented explicitly by these models, and their effects on the larger scales must be approximated in some way using only the large-scale information. Although steadily improving computing capacity continues to improve the dynamical forecast models through increased resolution, Palmer (2014a) has noted that hypothetically achieving cloud-scale (<1 km) resolution would require exascale computing, which in turn would require hundreds of megawatts of electrical power to run the computing machinery!

Even if all the relevant physics could somehow be included in atmospheric models, however, we still could not escape the uncertainty caused by what has come to be known as *dynamical chaos*. The modern study of this phenomenon was sparked by an atmospheric scientist (Lorenz, 1963), who also has provided a very readable introduction to the subject (Lorenz, 1993). Smith (2007) provides another very accessible introduction to dynamical chaos. Simply and roughly put, the time evolution of a nonlinear, deterministic dynamical system (e.g., the equations of atmospheric motion, and presumably also the atmosphere itself) depends very sensitively on the initial conditions of the system. If two realizations of such a system are started from only very slightly different initial conditions, their two time evolutions will eventually diverge markedly. Imagine that one of these realizations is the real atmosphere and that the other is a perfect mathematical model of the physics governing the atmosphere. Since the atmosphere is always incompletely observed, it will never be possible to start the mathematical model in exactly the same state as the real system. So even if a computational model of the atmosphere could be perfect, it would still be impossible to calculate what the real atmosphere will do indefinitely far into the future.

Since forecasts of future atmospheric behavior will always be uncertain, probabilistic methods will always be needed to describe adequately that behavior. Some in the field have appreciated this fact since at least the beginning of practically realizable dynamical weather forecasting. For example, Eady (1951) observed that "forecasting is necessarily a branch of statistical physics in its widest sense: both our questions and answers must be expressed in terms of probabilities." Lewis (2005) nicely traces the history of probabilistic thinking in dynamical atmospheric prediction. The realization that the atmosphere exhibits chaotic dynamics has ended the dream of perfect (uncertainty-free) weather forecasts that formed the philosophical basis for much of 20th-century meteorology (an account of this history and scientific culture is provided by Friedman, 1989). Jointly, chaotic dynamics and the unavoidable errors in

mathematical representations of the atmosphere imply that "all meteorological prediction problems, from weather forecasting to climate-change projection, are essentially probabilistic" (Palmer, 2001). Whether or not the atmosphere is fundamentally a random system, for most practical purposes it might as well be (e.g., Smith, 2007).

Finally, it is worth noting that randomness is not a state of complete unpredictability, or "no information," as is sometimes thought. Atmospheric predictability is typically defined with respect to the degree of statistical relatedness between forecasts and subsequent outcomes, characterized in terms of their probability distributions (DelSole and Tippett, 2018). A random process is not fully and precisely predictable or determinable, but may well be partially so.

To illustrate, the amount of precipitation that will occur tomorrow where you live is a random quantity, not known to you today. However, a simple statistical analysis of climatological precipitation records at your location would yield relative frequencies of past precipitation amounts providing substantially more information about tomorrow's precipitation at your location than I have as I sit writing this sentence. A still less uncertain idea of tomorrow's rain might be available to you in the form of a weather prediction that quantifies the uncertainty for the possible rainfall amounts in terms of probabilities. Uncertainty relates to how well something is known. Reducing uncertainty about random meteorological events is the purpose of weather forecasts, and reducing uncertainty about the nature of underlying natural phenomena is the purpose of much of scientific research.