Meteorology 5340
Environmental Programming and Statistics

Contents

1 Introduction

a. *What this course is about*

As summarized in the course description, environmental fields are overwhelmed with information, but methods are available to help acquire, analyze, visualize, and interpret the associated time series and multidimensional fields. This course is a compromise to allow more hands-on programming related to analyzing environmental data. *Statistics* as used in this course can be viewed as a part of *Data Science*, which encompasses both computer programming and statistics, among many other subjects. Data science involves what environmental scientists do: collect, prepare, analyze, manage, visualize, and store large volumes of information.

Knowing statistical methods is pointless if you don't know how to access and visualize the resulting information. Being a whiz computer programmer is not useful if you can't in the end find a use for what you are doing. It is frightening how often researchers don't understand statistical methods but apply them to large volumes of data and end up with physically implausible results. Just because a computer program ran and did not generate any errors, that doesn't guarantee what you found makes sense physically.

The goal of this course is to review and apply only a small number of core methods to examine data that may help you reach conclusions about environmental issues. Actually, good research usually ends up raising more questions than reaching solid conclusions. We will rely on the University's Center for High Performance Computing (CHPC) **Open OnDemand** web portal that allows you to learn basic programming concepts without getting too hung up on the nitty-gritty details. The **Python** programming language is used within the **Linux** programming environment.

The format of the course involves introducing concepts using Teams and leaving you time to work in class on assignments using the OnDemand web portal. My hope is that data science concepts and skills that you are exposed to here will be useful for next summer's project and in your classrooms.

These notes are the "statistics" text. It is necessary that you have a copy of the Second Edition of *Python Programming and Visualization for Scientists* by Alex De Caria and Grant Petty as a resource and reference to programming syntax. The objective is not to have you become an expert programmer, but help you to navigate a bit beyond Excel-type approaches with which you might be more familiar.

b. *Effective Research*

You may have some uneasiness regarding the applications of statistics to everyday life. Nearly every day some statistical study is reported in the media that is construed to prove the value of one substance or approach vs. another. A common public perception of statistics can be summarized as: statistics is useful if it confirms your biases and not useful if it requires you to question your beliefs. For example, consider the often misattributed quote: there are three types of lies- lies, damn lies, and statistics. But, let's be real- you rely on statistics all the time to *describe* events or *compare* outcomes (ERA, GPA), *infer* what's going to happen in the future

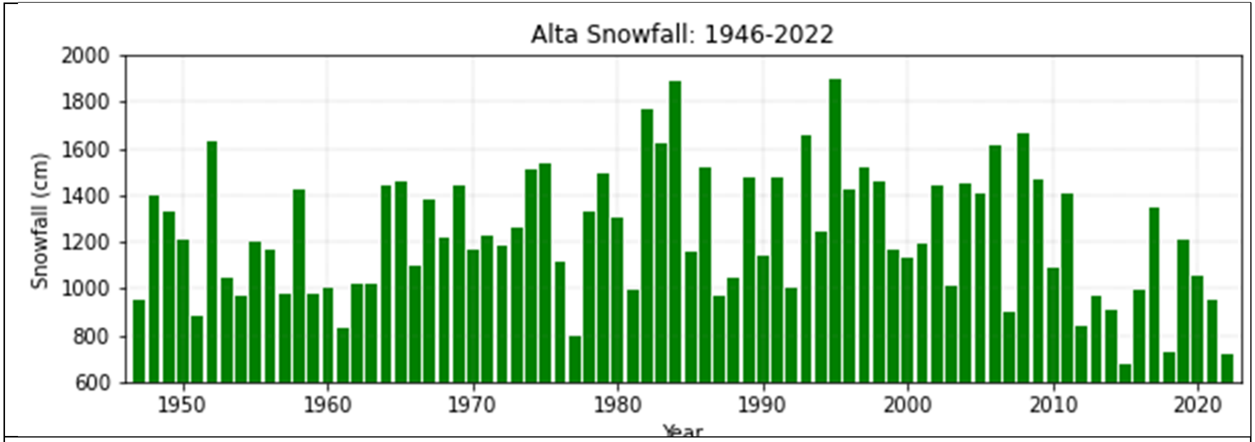

Figure 1.1. Seasonal total snowfall (cm) from November-April at Alta, UT.

based on past or current samples (polling for elections), *assessing risk* (is smoking bad for you?), and, what scientific research is all about, *identify* relationships within a large volume of data.

One of the goals of this course will be to emphasize that evaluating data requires you to check results and conclusions carefully. Could there be a problem with the data or do you have a programming error? What assumptions were made along the way to distill the data that might bias the results? Was the analysis approach really appropriate? What alternative explanations could there be? How statistically, as well as practically, significant are the results?

The bottom line is KISS- Keep It Simple Stupid. Most complicated programming and statistical approaches are just that- complicated. Machine learning sounds cool, and it is and increasingly easy to use. But, most people don't understand the assumptions and how and what those techniques are doing. My general rule is that if you have a really useful result, you should be able to see the glimmer of it in the raw data. The trick is to find the gold nugget in the gravel. Subjective evaluation of data is critical- if you have to manipulate and massage the data and then use some complicated analysis technique to distill the results, how useful will your results be? It may be useful if you are attempting to test and verify a well-defined hypothesis, but it may not have a direct practical application.

Abuses of statistics result from the common situation that if the only tool in your tool belt is a hammer, everything starts to look like a nail. Some statistical measures work better than others and some programming techniques are more efficient than others. A lot will depend on the goal

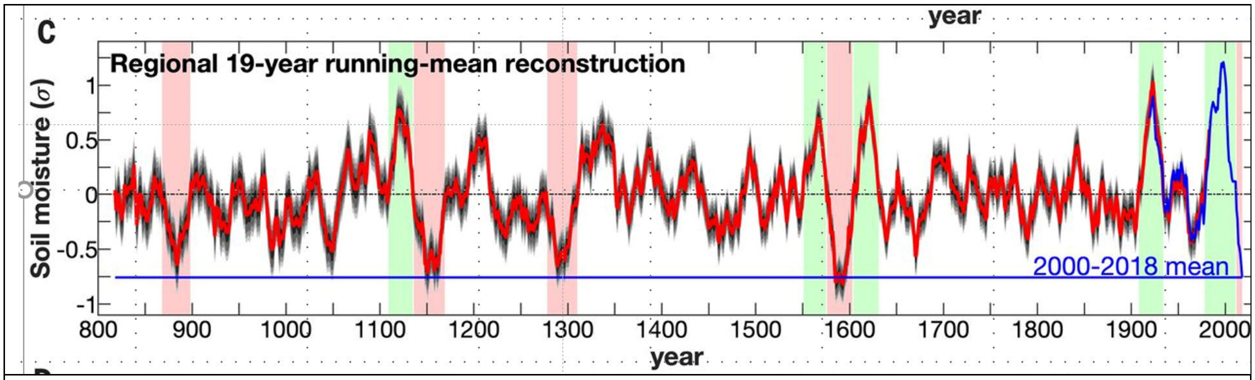

Figure 1.2. Summer soil-moisture reconstruction for the western U.S. highlighting recent megadroughts (Williams et al. 2020).

of the research and the type of data you are looking at. A goal of this course is to add a few tools to your tool belt- it will be up to you to figure out when is the right time to use them.

As introductory examples of environmental data, consider the time series of measurements of seasonal snowfall at Alta, UT from 1946 to the present in Fig. 1.1 and reconstruction of summer soil moisture for the western U.S. (Fig. 1.2) that is derived from tree-ring reconstructions. Rather than just accepting these records as representing climate variability from year-to year or on longer time scales, how were they created? What data were used? Who made the measurements and why? Snow doesn't really fall over a season- it happens during 1-3 day storms, so are seasonal totals the best metric to use? Yikes, the snow total this past winter was the second lowest from 1946 to the present.

In Fig. 1.2, the conditions during the past 20 years are similar to some of the most serious megadroughts in the past 1200 years. There are many questions we need to ask before assuming this index is a useful estimate of climate behavior for this region. How was it constructed in terms of the physiologic response of trees to climate? How sensitive is the methodology to local weather conditions and are the locales of the trees available for analysis representative of climate conditions regionally? What corroboration is available during the instrumental record that can be used to assess its value? What other factors do you think should be considered? Answering some of these questions will come from reading the paper, the supplemental material provided to support the paper, and their correction to the paper provided late last year.

Effective research requires preparation. Environmental problems are complicated. It is a bit unrealistic to expect that a phenomenon that evolves in time and space with complex interactions between variables will have a single causal factor that is completely apparent through a quick analysis of data. Rare events that have occurred recently are not due solely to anthropogenic forcing- how do we deal with multiple factors that might not be constant in time? That is a particularly vexing situation these days.

There are several basic steps to effective research:

- **distill** a general interest in a subject into **a specific question/hypothesis** that can be evaluated. While we are all likely interested in the impacts of global warming, it is a bit unrealistic to expect to examine all of the general circulation model data or long records of weather data in one study. Sometimes it is necessary to limit a study to what can be determined given the existing model or data assets. As your MSSST research progresses, you may need to search for other data that can help address the question of interest.
- **organize the data**. This can be the most tedious step, but the most critical. What data do you need? How should gaps of missing data be treated? What are the possible limitations of the data? Have you looked at the raw data enough to recognize what might be physically implausible values? Does the data make sense physically? Do you need to remove large signals that are not of interest to your question, i.e., seasonal or diurnal cycles, or model biases? You can avoid a lot of missteps by careful data preprocessing.
- **find relationship(s) among the data**. Your choices of analysis approaches are critical. Given your hypothesis and your data, what is the best way to proceed? It isn't going to be easy- as you proceed, you may end up having to reevaluate the goal of your research, find other data assets, or try other ways to analyze the data.

- **examine the significance of your results**. There is statistical significance, which you need to address, and then there is practical significance. Statistical significance is important- all too often people over fit data and then their results will be useless when applied to an independent data set. However, learn to recognize practical significance. Do your results make sense physically? Have you considered the amount of spatial and temporal dependence in the underlying data set? That is, the variations in environmental observations at one location may not be that different from those nearby. And, observations at one time may not be that different from those before or after the time of interest. Does the relationship hold up in an independent data set? While your results may pass a statistical significant test, are they of any use?
- **review thoroughly what you have done and document your analysis and results.** Did you cut any corners during the data preparation or analysis? The scientific ethos requires that based on the information you provide, someone else can take the same data and derive the same results. It is now often required in order to publish a scientific study that the data used in that study must be made available to others to evaluate your results.
- **submit your results and study for independent evaluation.** Peer review is critical for providing an independent opinion of the merits of the analysis that you have completed. Accepting criticism of one's work can be difficult. Be wary of any study that has not been subjected to peer review. If a project does not require peer review, then ask colleagues for an honest appraisal of the study's results.

Be aware, however, that refereed statistical studies may have serious flaws. Recognize that there is a *publication bias* to report positive results, no matter how inconsequential, as opposed to negative results. The influence of El Nino on the snowpack in the Wasatch is weak at best, and there have been many papers attempting to explain those weak relationships. However, it would be virtually impossible to have a journal publish a paper with the primary conclusion being there is no relationship between El Nino and Wasatch snowpack.

C. *Uncertainty*

Uncertainty is at the core of this course. Uncertainty arises because:

1. we can never measure the environment with complete accuracy and precision,
2. the environment is a chaotic system, which is a maddening combination of randomness and order arising from the characteristics of a complex nonlinear system,
3. our understanding of the environmental system is imperfect, so physical (and certainly statistical) models do not capture the complete behavior of the system.

Always assume that ANY observations are uncertain. Was a human observer involved? Did the same observer take the observations each time? What automated equipment was used? What metadata (data describing the data) are available to explain how the observations were taken? And, any chance the data were mucked up during some stage of the data processing before you received it- could programming errors have crept in?

I'm a pessimist by nature. If someone asks about some unusual observation that they saw in MesoWest, the first thing I do is begin thinking about why that observation is likely messed up.

Rarely do we know what the “true” state of the environment should be. An instrument can be calibrated to a standard in the laboratory, for example, distilled water in an ice bath should have a temperature of 0ºC (but what about impurities, dirt in the container, etc.?). However, I’m also more of an optimist than many- we can make sense of data even when imperfections in the data exist. If a wind sensor in a slot canyon only blows from the west or east, you may still be able to determine when a front passes that station in most cases, even with the biases inherent in the data.

We need to distinguish between the usually unknown (outside of the laboratory) “truth” and actual measurements.

- True value- value of a quantity sought through measurement, but unknown usually in the field

All measurements from instruments will provide an estimate of the true value, with varying degrees of success. A number of measures are available to gauge the uncertainty of observations:

- Accuracy- difference in response between a standard and instrument in varying environmental conditions
    - a measure of how close a measurement is to the “true” value
    - high accuracy can be expensive
- Limitation- capability of an instrument to give accurate readings within a specific range. At extreme ranges, readings may be less accurate and hence, limited
- Precision- how well repeated measurements of some quantity agree with each other.
    - a precise instrument can be inaccurate

Consider the following shots at a “bulls eye”, where the goal is to hit the X, the true value. While it is certainly best to have observations that are both accurate and precise, it may be too costly to do so.

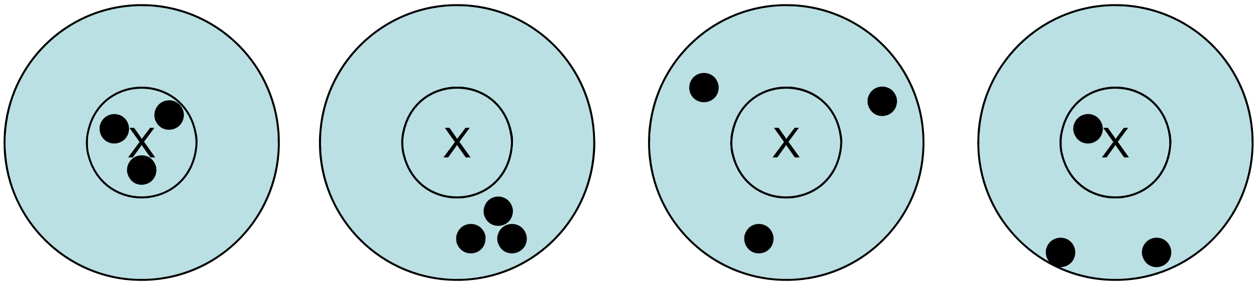

Figure 1.3. X is the “truth” while the filled circles reflect observations.

High accuracy	**Low accuracy**	**High accuracy**	**Low accuracy**
High precision	**High precision**	**Low precision**	**Low precision**
Small uncertainty	**Large uncertainty**	**Large uncertainty**	**Large uncertainty**

So, high precision cannot make up for inaccurate information (“*garbage in, garbage out*”). Many people end up confused when they hear the same wrong information from multiple sources (which often originate from a single highly unreliable source). *Judgment and integrity are critical* in all facets of life, particularly when evaluating the credibility of results obtained from environmental data.

The tendency in environmental fields is to overstate the amount of random error and underestimate systematic errors (or biases).

- Random- that which is not precisely predictable or determinable. Random errors arise physically due to sudden changes in the environment, due to turbulence or other processes. Random errors can also occur due to faulty equipment or observer carelessness.
- Systematic- errors arising from a consistent response of a measuring device to environmental conditions or faulty characteristics of instrumentation that occurs frequently. Systematic errors can vary as a function of weather regime: nonaspirated thermistors (thermometers over which there is no air blowing past mechanically) will tend to be affected by radiational cooling and heating when the winds are light, for example.

Look at the target plots above. High accuracy and low precision observations are an example of random errors while low accuracy and high precision observations are an example of systematic errors.

d. ***Population vs. Sample***

In the same way that we rarely know what the "true" value of a variable should be, we never know the entire population of true values as the environmental conditions change in time or space. We hope that we choose a sample of observations for analysis such that each element in the population has an equal chance to be selected, that is our sample is *representative* of the larger (usually unknown population). For example, we don't know what the future values will be, so there are clear problems with any sample we choose. For example, the increasing trend of atmospheric carbon dioxide concentrations implies that a sample taken from the recent past can not reflect accurately the population of carbon dioxide concentrations that includes future values. When we choose a sample, we must try to avoid *selection bias* (cherry picking the environmental situations we study).

Also, because of the serial dependence of environmental data (observations collected consecutively will tend to be similar to one another depending on the time scale of the phenomenon being measured), it is often difficult to have each element of the sample be representative of the span of the observations possible in the population. Further, numerical model errors are often such that samples from a model tend to be less variable than observed samples, so that a sample derived from model fields will not be representative of the observed population.

Selecting the sample for analysis is a critical aspect of organizing the data and depends on the question to be addressed by the study. If we want to examine the frequency of occurrence of major snowstorms at Salt Lake City, does it make sense to include data from the entire year or limit the analysis to data from the cool season only?

An obvious rule of thumb is that your sample should be large enough to capture the phenomenon of interest many times. **"Degrees of freedom"** refers to the number of independent elements in the sample; the number of degrees of freedom is usually much smaller than the total number of members in the sample in environmental data sets. You may want to use only a fraction of the total data available to begin your analysis. Then, the remaining data will be an independent sample that you can use to evaluate and confirm your results. Alternatively, procedures can be used to rerun the analysis hundreds of times omitting randomly data each time in order to develop confidence in the results. All too often, people assume that their sample is drawn randomly from the population, when in reality, their sample grossly underestimates the variability inherent in the population.

To illustrate these points, consider Fig. 1.4. Assume that we are measuring some phenomenon for which the population consists of only 1 possible "true" value equal to -1. So the population mean μ (average of all of the true values) in this case is the same as each individual true value and equal to -1. Then, we make a total of "n" repeated measurements (in this case a million), which is our sample. One of those measured values, x_i equals 1 so the measurement error for that specific case is 2. The histogram counts all million values into bins. It is more likely in this made up example that we measure values near 0 and never measure values outside the lower and upper limits. Then, the sample mean is

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

In our example, the sample mean is 0, so we have a systematic error or bias of +1. If the width of the bell curve narrowed, the precision of the sample would increase. As the mean of the sample shifts towards -1, then the accuracy of the sample would increase.

It is also important to recognize when events may really be independent of one another. The probability of getting heads when you flip an unbiased coin is always 50%, no matter if you have had 10 heads in a row before. Many times people confuse streaks, or "hot hands" as being real when they are not. Casinos stay in business because gamblers' perceptions differ from reality- a person's intuitive misunderstanding of causality (I'm wearing my lucky shoes) departs from well-founded odds of a likely (or unlikely) outcome. And, clusters of events do happen by chance- two people could be winning at the same blackjack table *by chance,* not because it is a lucky table. Specific, rare health complications might occur in one town at a higher rate than "normal" *by chance*, independent of local environmental conditions, simply because lots of people live in lots of towns.

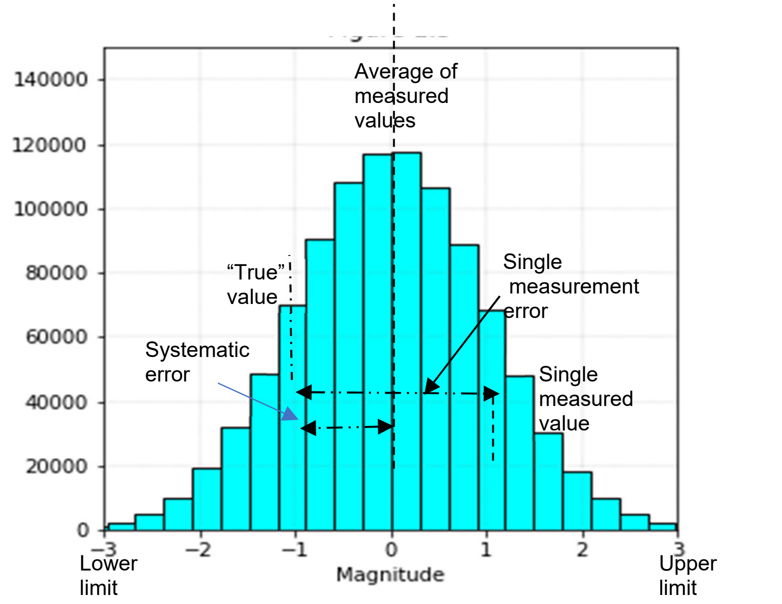

Figure 1.4. A million member sample. The True value is assumed to be -1. The systematic (average) error would be +1.

It is also important to clarify different types of samples within a population. If two samples are not drawn from the same portions of the population, then they are **independent** of one another. For example, a time series of values from 2000-2009 would be independent of a similar time series from 2010-2019. Two samples are **dependent** if the values in one sample are used to determine the values of the other sample. If we try to estimate the snowfall amounts each year at Alta from the snowfall at Salt Lake City, then those Alta estimates are dependent on those at Salt Lake City.

In **machine learning** (a fancy term for using a wide class of statistical tools), a **training** dependent sample is used to determine *parameters* required to estimate some quantity from other data. Then, those *parameters* are evaluated by using an independent **validation** sample. Finally, another independent data sample, the **test** or **hold-out** sample, is used to evaluate the effectiveness of the technique to estimate the quantities that are desired.

e. ***Reducing Dimensionality***

Programming and statistical methods applied to environmental data typically involves reducing the dimensionality of the data to a manageable size. Which variable(s) do we need to consider? Can we consider one variable (univariate analysis) or must we consider multiple variables (multivariate analysis). What time scales are we interested in? Hours, days, months, years? And, what region (local, regional, globally) or level in the vertical (surface, subsurface, upper air)? Are the data available on a spatial grid or at specific points? Reducing dimensionality pops up as all kinds of fancier terms: neural networks; principal component analysis, etc.

Statistics often is misinterpreted as bookkeeping. What is the warmest temperature on record at Salt Lake City? What is the biggest snow storm at Alta? To even begin to answer those questions raises a number of issues. What period of record are we considering? What instrumentation? What about siting issues: where are the observations located (airport, downtown, Alta, etc.)?

We do need to distinguish between weather and climate:

- weather- state of the atmosphere and, more broadly, state of the environment
- climate- aggregate summary of the weather, or, more broadly, aggregate summary of the environment

How we go about aggregating the sequence of states of the environment is a critical issue. Much of the present climate framework in the U.S. is tied to somewhat outdated practices: climate normals are defined in terms of 30-year summaries that shift over time. The recent change in the definition of the normal period has been big news of late with the shift from the 1980-2010 period to the 1990-2020 period.. “Normal” refers to the mean, the average of the values observed over time. It is designed to tell someone what the typical weather for a location might be. But, we’ll see later that the mean is not always a good measure of the typical weather- it is sensitive to occasional outliers (either real ones or ones that arise from failing to identify erroneous observations).

Focusing on extremes (for example, record highs and lows) can also be misleading. Is it critical if the temperature drops in a few minute period to some really low value, or is it more important if it stays at a slightly higher value, but still unusually low, over a several hour period? Many of our current practices of recording mean and extreme conditions is simply an attempt to reduce the dimensionality of the volume of data available. However, with the sorts of computing resources now available, we should not feel so constrained to follow such practices, and, at the least, be wary of how those practices may constrain our understanding of the way the environment works.

Statistical approaches are useful for weather and climate because:

- both are controlled by innumerable factors, which we hope to segregate into a few critical factors from the rest that, for the most part, simply contribute to background noise
- the characteristics of the system include linearly unstable processes that cause growth of small features into larger ones (a topic covered in other atmospheric science classes)
- the characteristics of the system (dynamics, thermodynamics) are nonlinear and include discrete step functions (i.e., rain/no rain) that can lead to the amplification of small errors into large ones
- the system is dissipative, which implies a tendency for “stationarity”, i.e., the climate system will remain stable and not run away completely from the current state. Global warming is not going to cause earth to turn into Venus, for example.

So, if we want to predict some future outcome, what variables must we consider? What locations? Over what time interval? Dealing with location (horizontal and vertical), time, and variable simultaneously is best left to numerical models of the entire climate system. The nonlinearities and instabilities make the environment unpredictable after characteristic times that differ among the various subcomponents (atmosphere, ocean, ice, etc.). The combination of noise and damping in the environmental system makes statistical predictive approaches credible. Deterministic approaches to forecasting future states of the environment (i.e., that the system is known and predictable at short lead times once the initial state is specified) may be less accurate than statistical approaches, which typically presume that there are a range of likely outcomes given the uncertainty inherent in the system.

f. ***Descriptive vs. Inferential Statistics***

The distinction between descriptive and inferential statistics ties together many of the above points:

- Descriptive- organization and summarization of data, which may include using statistical models to interpret the volumes of data

- Inferential- figuring out why the environmental system behaves the way it does using data.

During your career you may be faced with the task to extract information from environmental data. It is unlikely you will develop a new coupled model of the earth-atmosphere system from scratch. Many of you will likely be handed chunks of environmental observational data that has been collected. An effective evaluation of data entails the use of statistics descriptively and inferentially. The process of data analysis requires brute force detective work: being organized and thorough as well as following through and experimenting with different approaches to examine the data. It also requires insight and intuition into the workings of the environment: you must understand the goals of your investigation and the flexibility to adapt as your understanding improves during preliminary analyses. And, you want to use appropriate programming methods so others can use what you have written.

Statistical inference is the goal- we want to draw meaningful conclusions from the data. That means we need to have a plan and an idea about what to expect when we analyze the data. We need hypotheses. However, *statistical analyses can't prove much- but, data analysis can rule out incorrect hypotheses*.

In court, you are expected to be assumed initially to be innocent (not guilty). That is the *null hypothesis*, you are not guilty. The prosecutor's job is to get the jury to reject that you are innocent and provide an alternative hypothesis that you are guilty beyond a reasonable doubt. The defending attorney's job is not to prove you are innocent, all that is necessary is to come up with other alternative hypotheses, such as someone else did it or raise the uncertainty about your level of guilt above reasonable doubt. We will see that "beyond a reasonable doubt" for a statistical inference is related to how often we might expect something might happen by chance- once in 20 cases, once in one hundred cases, etc.

We have to be very careful about how we infer a meaningful conclusion. If we want to be really careful and only reject the null hypothesis at a high level of certainty (avoid false positives by saying it can only happen once every ten thousand cases), then there are going to be more times when we should have rejected the null hypothesis and didn't (false negatives). In other words, guilty people will go free more often to insure that no innocent person is sent to jail. Perhaps, a less threatening example is designing a spam filter. Your spam filter begins by assuming every email is not spam. You should be resigned to let through a few unwanted emails to avoid blocking an important email. Right now the University's spam filter is labeling some critical emails sent to me as spam- that is not good programming.

g. *Replicability and Reproducibility*

Replicability is assessed by researchers who perform an experiment under exactly the same conditions multiple times. Replicability reflects the technical stringency or precision of a specific experiment. Thirty years ago, two researchers at the University of Utah put out a press release claiming they managed to and replicated generating excess heat at room temperature: *cold fusion*. That generated a huge hubbub, but it was not reproducible by others to the extent of their claims. **Reproducibility** is the extent to which measurements or observations agree when performed by multiple researchers.

The number of journal articles published worldwide is now over 2 million annually. A quarter of these are in the field of biomedicine of which a large fraction of those are never even referenced by other studies. Is it any surprise that many articles are being published that contain irreproducible results? Search technologies are helping identify issues, but only about .4% of articles are retracted of late. Retractions arise from plagiarism, data manipulation (especially in figures), and proven data falsification.

Reproducibility, rigor, transparency and independent verification are cornerstones of the scientific method. Just because a result is reproducible does not make it right (or useful). *Reproducibility* is assessed by performing similar, but not identical, experiments at different times, in different locations, and under somewhat different experimental conditions. *Replicability* reflects the technical stringency or precision of a specific experiment. A precisely conducted experiment can be inaccurate, and an accurate experiment may be performed imprecisely—especially in biomedicine where many factors can account for irreproducible results.

Nassim Taleb highlighted a fundamental flaw of traditional statistical methods and the concept of reproducible results using the term **black swan event (**https://en.wikipedia.org/wiki/The_Black_Swan: The_Impact_of_the_Highly_Improbable). The rarity of black swans in nature lends itself to the concept of rare and unpredictable events (at least unpredictable to people who are not experts on swans). Scientists are not immune to explaining such events, retrospectively (*a posteriori*) rather than in advance (*a priori*). Hence, statistical approaches might be in theory reproducible for most environment conditions, but not for black swan events. Because of their inherent rarity, it is difficult to explain away an irreproducible result on the basis of a rare, singular event for which no other evidence is available. Handling black swan events introduces the concepts of vulnerability to rare events and the need for resiliency to cope with such events. To most of us, the COVID-19 situation has been a black swan event. But, public health experts certainly understood such events could happen and many expected something of this sort was predictable and we should have been better prepared for it.

Guidelines from major journal publishers now recommend that journals include in their information for authors their policies for statistical analysis and how they review the statistical accuracy of work under consideration. Errors in design, analysis and interpretation of data science/statistical approaches could be inadvertent or intentional. Bias clearly plays an important role in promoting false-positive results. Assumptions regarding random and systematic errors contribute as well. Reasons for such bias include:

- lack of experimental balance leading to an impassioned belief in one particular experimental outcome clouding objectivity;
- perceived pressure to publish for academic advancement or to enhance the likelihood of competing successfully for grant funding;
- the lack of appeal to publish negative (or neutral) results in most high-impact journals.

h. ***Putting Statistics to Work***

We will be using the Open OnDemand framework to alleviate the drudgery associated with learning Python language syntax and statistical methods. You must have the Second Edition of *Python Programming and Visualization for Scientists* by Alex DeCaria and Grant Perry as a reference. The objective of this course is to make you aware of programming methods and foster the ability to work both independently and collaboratively to evaluate information. Don't expect us to solve all of your programming mistakes. You will have the opportunity to use Teams to communicate with others and bounce ideas and solutions off one another.

It is important that you first understand what you need to do to complete the course assignments. Then, take advantage of Python to relieve the tedium of doing the statistical calculations. That may often require you to do some of those calculations manually and then make sure and verify what you get using the Python notebooks and programs.

Share code features with others (and us!), if you do so. *However, using someone else's code is plagiarism. If you work in a group, make the final product yours, not someone else's code copied verbatim.*

As with all programming, there are good ways to approach an assignment, as well as very frustrating approaches. Here's a few reminders:

1. understand what you are expected to do. Ask questions and get clarification before trying to write code.
2. look at the example code, run it interactively, and understand what it does. You won't be expected to start from scratch. Again, ask questions if something is not clear and pay close attention to the details of the example code.
3. Read the notes and text and look online for techniques and tricks if you're stuck. There are lots of resources to solve problems.
4. Don't assume there is only one way to do something, but, recognize I am generally expecting a particular approach that csould be used. Let me know if you've found an alternative approach that works.

Programming is an iterative process. You do something, it doesn't work, you make a change, and it still doesn't work. Then what? First, when did it stop working? Did the example code do what it was supposed to do? Then, what did you change? Have you looked at the variables? Did you look at some intermediate values? Don't assume that just because you didn't get an error message that everything is coded correctly. Check your results. Do they make sense? Look very carefully at what you are doing and then ask others in the class if you are still struggling.

i. ***Navigating Online, Communicating and using your own computing resources***

The syllabus and assignments are online in Canvas. Canvas is good to handle assignments and know what is due and when. Codes will be available via OpenOnDemand. You must have the example codes and class notes at your fingertips. You will be able to do all of the work using your own computing resources as long as you have decent access to the internet.

I will be using Teams to provide quick updates to everyone or chat individually with you. The Teams workspace for this class is: Teams. You will receive an invite before the class starts about that.

2 Exploratory Univariate Data Analysis

Environmental fields are awash with data. The first priority of any analysis is to simply spend time looking at the data in a variety of ways. Then, the next step is to reduce the dimensionality of the data sample by summarizing the data. Different types of data lend themselves to different approaches, so use examples from other studies or papers to gain ideas as to what might work for a particular data set. We'll begin by examining data samples (level of the Great Salt Lake and temperature and precipitation summarized for the state of Utah as a whole) using univariate techniques (i.e., the analysis of one variable is assumed to be independent of any other variables). Many of the concepts described here you may have already used extensively. The objective is to increase your awareness of other tools that may be more appropriate, particularly given the limitations of some commonly used techniques.

Programming skills are developed by seeing how others do it. Sometimes you learn as much from poorly written code (what you might see in this class!) as well as excellent code. There are certainly different ways to examine time series using univariate approaches and different ways using Python to do so. It will be very important for you to pay close attention to the Python code corresponding to the concepts in this chapter. It doesn't all have to make sense yet- we will be explaining more about the nuts and bolts of plotting later, for example. Use the text to look up concepts that we may not have presented to you yet. The files we will use for the the class are accessible via https://home.chpc.utah.edu/~u0035056/atmos_5340/. You will use the files in the subdirectory data/gsl_yr.csv, data/utah_precip.csv, and data/utah_temp.csv and python notebooks in the subdirectory chapter2/ (chapter_2_2022.ipynb and chapter_2e_2022.ipynb).

a. Examining Time Series of Data

Geophysical data are usually collected sequentially, often at regularly spaced intervals. However, the data collection interval may change during the period of record, which introduces issues as far as how to summarize the data. Let's begin with the record of the level of the Great Salt Lake as a function of year from 1895 to 2021. Data are available from https://waterdata.usgs.gov/nwis/uv?site_no=10010000. Be sure to poke around that site to understand how and what the observation are. We'll also use estimates of annual precipitation and temperature for the state of Utah for the same period. These are available from http://www.wrcc.dri.edu/cgi-bin/divplot1_form.pl?4203.

First, look at the raw data files that you downloaded. The middle column of the lake level file is the number of observations. During the early years, the number of observations is only a couple per year. Recently, daily values of lake level are available (and actually, they are now 4 times per hour). Hence, there may be some uncertainty about the lake level in the early part of the record relative to the latter part simply on the basis of the methods used to record the observations. (However, we'll see that the large serial dependence of lake level, i.e., that the lake level varies slowly, mitigates this problem to a large extent.)

The annual precipitation and temperature are derived from Cooperative Observer reports, summarized into climate divisions, and then aggregated into the statewide average. There's a large number of steps and assumptions behind those calculations, so don't simply assume that these annual estimates are really representative of the state as a whole.

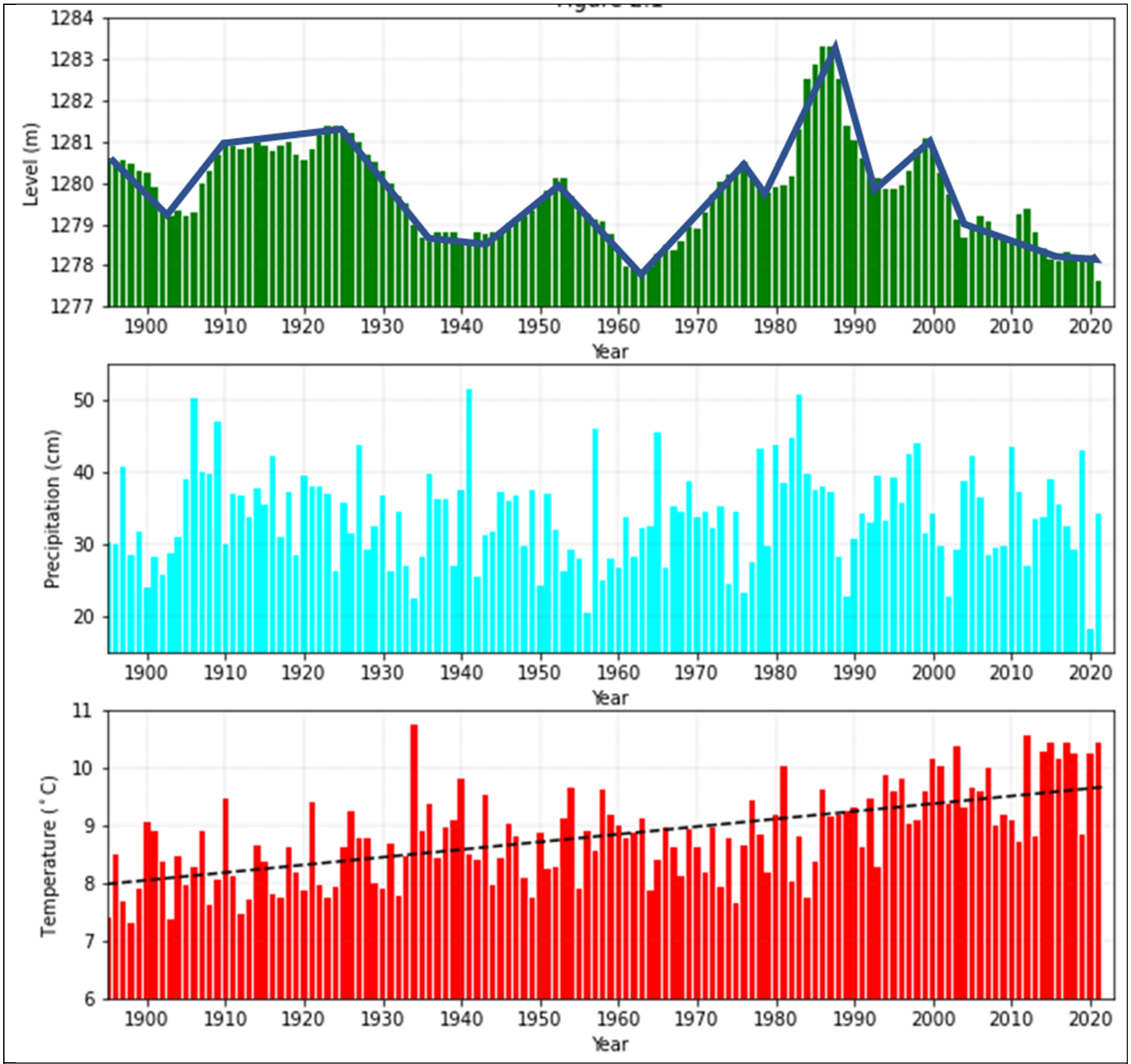

Figure 2.1. Top panel. Level (green bars, m) of the Great Salt Lake. Middle panel. Annually averaged precipitation (cm) in Utah. Bottom panel. Annually averaged temperature (°C) in Utah. The heavy lines in the top panel approximate how many unique periods exist during the record for lake level.

Code is provided to read, analyze, and plot the following figures using Python. The data are read into arrays and then time series are plotted as **bar plots**. As shown in the top panel of Fig. 2.1, the lowest lake levels were observed recently and in the 1960's while the highest water years were in the 1980's. The trend during the past decade has been for lake level to be dropping and then remaining relatively flat. The serial dependence of the lake level data is evident, i.e., the value of lake level in one year is usually similar to that in adjacent years. A simple way to estimate the number of independent values in a sample is to draw subjectively line segments that reproduce the primary features of a time series as shown by the heavy line in the top panel of

Fig. 2.1. Then, the **degrees of freedom** is the mean value plus the number of line segments (or count the points required to draw the lines), which is ~20 in this case Hence, even though there are 127 years in the sample, only about one in six of those values are independent of the others.

Ignoring serial dependence in time series is a really big failure in many geophysical statistical studies. For a time series comprised of from two to a million data values, it is entirely possible that there are as few as 2 independent values (degrees of freedom) in a time series. *If the time series has a very large trend, then it can be described entirely by the mean value plus the slope of the line.* All too often people assume minor "wiggles" in a time series are relevant when the statistical approach they are using weights heavily trends and other dominant features.

Let's look at the annual total precipitation for the state of Utah in the middle panel of Fig. 2.1. The wettest year for the state as a whole took place during 1941 and 2020 was the driest. There are clearly some strings of years with greater than usual precipitation as well as drought episodes. The string of wet years in the early 1980's corresponds to the increase in lake level of the Great Salt Lake, for example. I'd guestimate about 50 or more line segments would be required to reproduce the major features of the time series, which would suggest that the values roughly every 2-2.5 years are independent of the others. It's not really important to try to draw those line segments in complicated time series- there are techniques to define the amount of "persistence" in a time series as will be discussed later.

Now, let's examine the year-to-year changes in air temperature in the state of Utah. The temperature in Utah during the late 90's through the recent years have been warmer than that during any other decade during the past 100 years. Note that 2019 was cooler than other recent years. Temperature during 1934 appears pretty unusual and is the only year where the annually averaged temperature was above 10ºC before the 1990's. Annual temperature greater than 10 ºC have been common of late. The serial dependence from year to year is clearly less for temperature than for lake level, i.e., we have many more independent values in our sample of air temperature than we have for lake level. However, temperatures during the early part of the record (on or before the 1920's) certainly are lower than those of late.

The "**linear fit**" or trend over time for Utah's temperature is shown by the dashed black line. Temperatures in recent years are roughly 1.5°C higher than they were around 1900. How such a linear regression is derived will be discussed later. The question for you to ponder is whether this trend line adequately describes what has happened over the past 127 years? Has it been a steady rise incrementally from year to year? Or, are there important changes on shorter time scales than the full 127 years that we should consider?

b. Data Distributions and Histograms

An obvious first step when examining a data set is to order (sort) them from smallest to largest or vice versa. See the Python code for that. Take a moment and look at the resulting ordered data in the program. The first (last) element is the lowest (highest) value and equal to 1277.8 m (1283.3 m) for lake level. Without some formatting, the values in the Python program would have five decimal places - that does not indicate precision in the data. Look at the original data and note that the original values are in tenths of feet so it is reasonable to treat the data to tenths of meters.

Of course, computing the maxima, minima, and **range** (difference between the highest and lowest value) can be done easily. The Great Salt Lake has fluctuated in this sample over a range of 5.6 m. It is important to recognize that for a time series of environmental observations, the serial dependence of the data is lost when we sort by value. So, there is a tendency with sorted of data to overemphasize the total number values in a sample (127 years in our case) rather than how many independent values there may be (~20).

Histograms are a convenient way to summarize the sorted data by aggregating them into bins ordered from smallest to largest (Figure 2.2). The most basic rules of thumb are simply to choose bin widths for histograms that give a relatively smooth appearing histogram or subdivide the range into convenient subintervals for labelling. The lake level has been most commonly between 1279- 1281 m (upper left figure). If we divide it up into .5 m intervals (upper right panel) we see that values from 1278.5 to 1281 m are roughly equally likely to occur. There are clearly some outliers, with a few years with levels greater than 1282.5 m and no years in the sample with values between1281.5 and 1282.5 m.

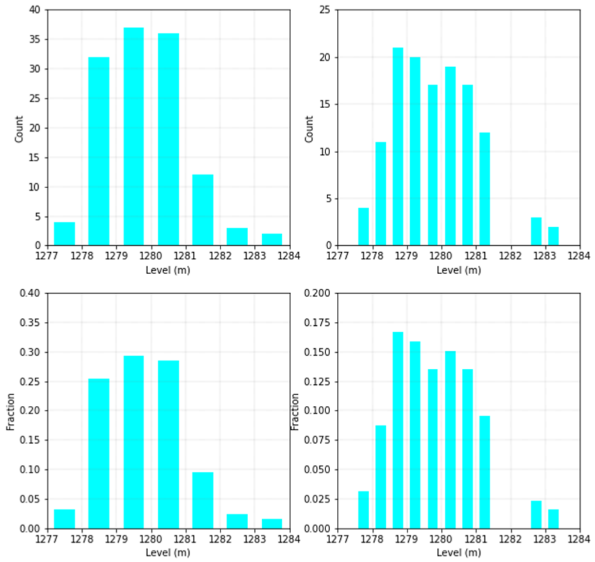

Figure 2.2. Histogram of lake level using 1m (upper left) and .5 m (upper right) intervals. Fraction of the total that lake level falls within 1 m intervals (lower left) and within 0.5 m interval (lower right).

The lower panels in Fig. 2.2 illustrate how the raw counts in each bin can be divided by the total number of years (127) to yield the fraction of the data in each bin. These are **empirical probabilities** of how often values lie within specific ranges. For example, ~60% of the time the lake level has been between 1279 - 1281 m.

If the percentage of the total contributed by each bin in the histogram is added from smallest to greatest, then the **cumulative frequency distribution** is created (Figure 2.3). Only 10% of the time the lake level has been above 1281 m, 50% of the time the lake level has been above ~1280 m, and 10% of the time the lake level has been below 1278.5 m. If the cumulative probability was computed from a coarse histogram (such as if it was done from the values in Figs. 2.2), then there would be distinct transitions in the cumulative probability from one discrete value of GSL level to another. However, the cumulative probabilities are computed incrementally in Fig. 2.3 for each data value, which is why there is the fine scale chatter in the distribution.

c. *Central Value, Spread, and Symmetry*

The characteristics of a sample of data are often summarized in terms of:

- **central value** (central tendency or typical value)
- **spread** (variation or dispersion about the central or typical value)
- **symmetry** (degree to which the values tend to be larger or smaller than the central value)

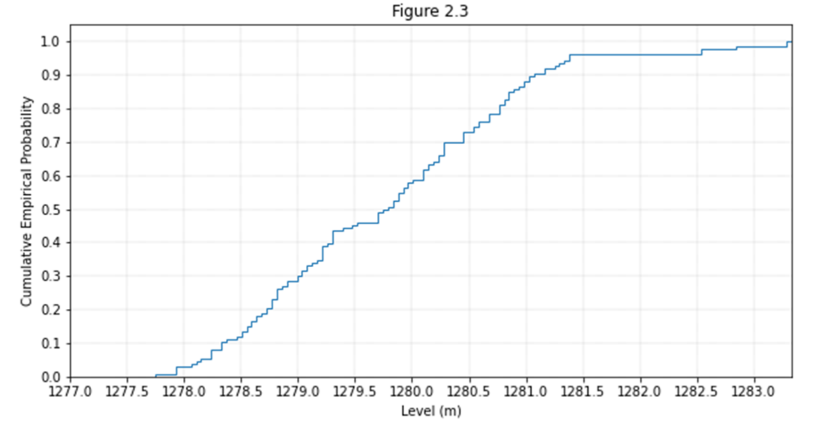

Figure 2.3. Cumulative frequency distribution of lake level.

These quantities are often referred to as the first, second, and third **moments** of the data. There are also higher moments: the fourth moment is **kurtosis** that evaluates the degree to which a sample has multiple peaks in its distribution or whether the distribution is relatively flat.

Whatever approaches we use to summarize the data, we want measures that are:

- ***robust***- which means not overly sensitive to the characteristics of the entire sample of data values. In other words, we want the measure to perform reasonably well no matter how the data values are distributed.
- ***resistant-*** not unduly influenced by outliers in the sample. For example, the range is not resistant to outliers.

Histograms and cumulative frequency distributions help to define visually the central tendency, spread, and symmetry of the sample. **Quantiles** are defined as percentage thresholds of the data that can be estimated visually from cumulative frequency distributions or computed easily.

- q_{25} – lower quartile- 25% of the sample lies below that value and 75% lies above
- q_{50} – median- 50% of the sample lies below that value and 50% lies above
- q_{75} – upper quartile- 75% of the sample lies below that value and 25% lies above

The **median** is a very good measure of the central tendency of the data, i.e., the typical value. It tends to be robust and resistant. Terciles (thirds) and deciles (tenths) get used frequently as well. If the sample is small, then it is easy to go through an ordered list and count off where the percentage thresholds will lie. If the quantile falls between two values, then the average of the two adjacent values is used (i.e., if the sample contains 4 values, then the median is the average of the second and third value). Experiment by using terciles and deciles as well. In our case, the lake level has been below 1281 m roughly 90% of the time while it has been below 1279 m roughly 30% of the time.

Box and whisker plots as in Figure 2.4 are a simple way to visualize the range, median, and quartiles of the data as well as outliers. The center 'notched' line is the median. The top of the box is the upper quartile, the bottom of the box is the lower quartile. The difference (q_{75} - q_{25}) between those two values is the **interquartile range** (IQR). The IQR is a robust and resistant measure of spread within the data set. The ‘whiskers’ are defined arbitrarily here as the 10th and 90th percentiles. Outlier values above and below those thresholds are shown as circles.

For all three variables in Fig. 2.4, the outliers above the 90th percentile tend to be more spread out, which we’ll see is an indication of ‘positive skewness’. In the case of the GSL level in the left panel, the high water years in the 80’s relative to the values observed in other years were unusual. Boxplots can be very useful when the data are noisy to quickly define values that may be physically implausible. However, extreme values such as those high water level years are not erroneous, they happened. Similarly, the precipitation values in 1941 and a couple of other years are really unusual based on the sample in the center box and whisker plot.

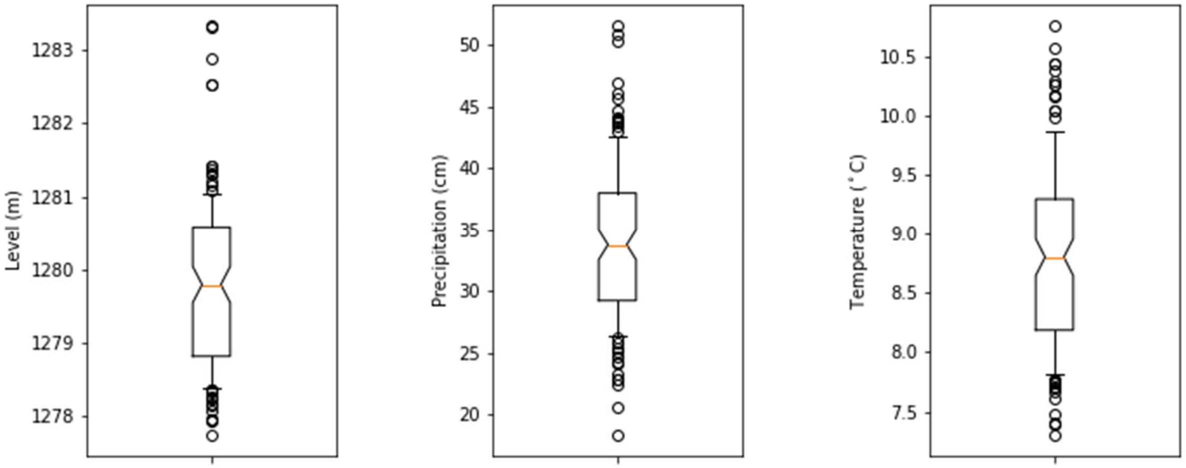

Figure 2.4. Boxplots of lake level (left), Utah precipitation (center), and Utah temperature (right).

Another central tendency metric is the **mode-** the most frequently occurring value. One way to identify the mode is to determine the center of the bin in a histogram with the largest number of counts, e.g.,1278.75 m for the lake level from the upper right panel in Fig. 2.2. The mode is a robust and reliant metric of central tendency. Computing the mode using the raw data can be misleading as shown in the python code- it is better to bin the data a bit before estimating the mode.

A traditional measure of the central tendency or typical value of the sample is the **mean**, which is not a robust and reliant measure of the central value:

- $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i$ *(2c.1)*

For convenience in the sample codes, the variables lev, temp, and ppt are loaded into columns of a new **“numpy”** variable with the unoriginal name ‘array’. These notes will continue to refer to the lake level variable alone, which is the first column of array. For the sample of lake levels, the mean and median are the same: 1279.8 m. But modify your data set by throwing in a bad value for the first element (e.g., 9999), which can commonly happen if the data file becomes corrupted for some reason. Obviously, the mean is now much higher, while the median remains unchanged.

You will see in the code that we begin to use the **"pandas"** module at this point too. Pandas provides many useful statistical functions and convenient ways to manipulate data in "**dataframes**".

The **trimmed mean** $\overline{X}_\alpha$ is a more resistant measure of central tendency, since a fraction of the high and low values are removed before the mean is calculated:

- $\overline{X_\alpha} = \frac{1}{n-2k}\sum_{i=k+n}^{n-k} x_i$ *(2c.2)*

The kth highest and lowest extreme values are removed before computing the mean from the rest of the sample.

Now let's look at measures of spread. You should already be aware that the maxima, minima, and range are not robust and resistant measures of spread. The **median absolute deviation** (MAD) is a more robust and resistant measure of spread and uses all of the data rather than the central core of the data as with the IQR.

- MAD = median | $x_i - q_{.5}$ |

MAD is computed by taking the absolute difference between each value and the sample median and then taking the median of that resulting sample.

MAD is .9 m for the GSL level. The median absolute deviation tends to be a conservative measure of spread.

The **standard deviation**, s, is a common measure of spread that is not resistant to outliers or robust. The square of the standard deviation is the **variance**, s^2, and is called an **unbiased estimate of the population variance** (and s is referred to as an unbiased estimate of the population standard deviation):

- $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$ *(2.c.3)*

For the GSL level, the standard deviation is 1.16 m and the variance is 1.34 m^2. Hence, the standard deviation provides an indication that there is roughly 1 meter of variation or dispersion about the central value of 1279.8 m and the MAD indicates there there is roughly 1 m of variation around the median of 1279.8 m. The MAD is lower because it is less influenced by those rare years of really high lake levels in the 1980's.

It is very important to pay attention to the units: the standard deviation has the same units as the quantity itself, while the variance has those units squared. Why is the variance s^2 calculated by dividing through by n-1 rather than n as we did when computing the mean? Consider a population with mean 0 and population variance σ^2. Then the variance of the population can be computed in the same manner as the population mean:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \overline{x^2} \quad \text{(2c.4)}$$

When we use a sample of n independent values (which may be a small sample), and if we compute:

$$s_x{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \quad \textit{(2c.5)}$$

then $s_x{}^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{2}{n}\sum_{i=1}^{n}(x_i \overline{x}) + \frac{1}{n}\sum_{i=1}^{n}(\overline{x})^2$

Now we need some summation identities:

$\sum_{i=1}^{n} a = na$ where a is a constant and $\sum_{i=1}^{n} ax_i = na\overline{x}$ *(2c.6)*

Then $s_x{}^2 = \overline{x^2} - \frac{2}{n}\overline{x}\sum_{i=1}^{n}(x_i) + \frac{1}{n}\overline{x}^2 n = \overline{x^2} - 2\overline{x}\overline{x} + \overline{x}^2 = \overline{x^2} - \overline{x}^2$ *(2c.7)*

Let's stop here a moment and recognize that the final relationship in 2c.7 is a convenient way to compute the variance that does not require removing the mean first and then summing. Instead, we can sum the squares of the values (first term) at the same time we are summing the values (second term) and later squaring the mean value. That approach is an old school way of computing the standard deviation and variance.

Now, if we had a very large sample, the second term $\overline{x}^2$ should be zero, since the population mean is zero. But, for a small sample, it may not. Given that the sample is supposed to be comprised of independent values, then it can be shown that:

$\overline{x}^2 = \frac{1}{n}\overline{x^2}$ (you'll see that this comes from the central value theorem later). As n gets big, then it will trend to zero. So, $s_x{}^2 = \overline{x^2} - \frac{1}{n}\overline{x^2} = \frac{n-1}{n}\overline{x^2} = \frac{n-1}{n}\sigma^2$

Comparing *(2c.3)* to the above, $s_x{}^2$ = (n-1)/n s^2, which is why s^2 is an unbiased estimate of the population variance; $s_x{}^2$ is the *sample variance*. Why should you care? It usually is not a big deal, as the differences between the sample and population standard deviation are likely small if the sample size is large (more than 50 or so). However, we'll see later that it can be important to differentiate between what we *measure* from a sample and what we *estimate* for the population.

Symmetry is a measure of the balance around the center value. Skewness (γ) is a nondimensional measure of asymmetry. If γ is close to zero, then the sample is close to a bell curve with roughly equal numbers of negative and positive outliers. Skewness will be negative, when data are spread more below the mean than above the mean and positive when there are more positive outliers than negative ones. Skewness is neither robust nor reliant.

- $\gamma = \dfrac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^3}{s^3}$

In the case of the GSL level, the skewness is .61, which indicates that there are larger departures above the mean, i.e., the large positive outliers "skew" the distribution of lake level. The annual precipitation and temperature are also positively, but less, skewed than lake level.

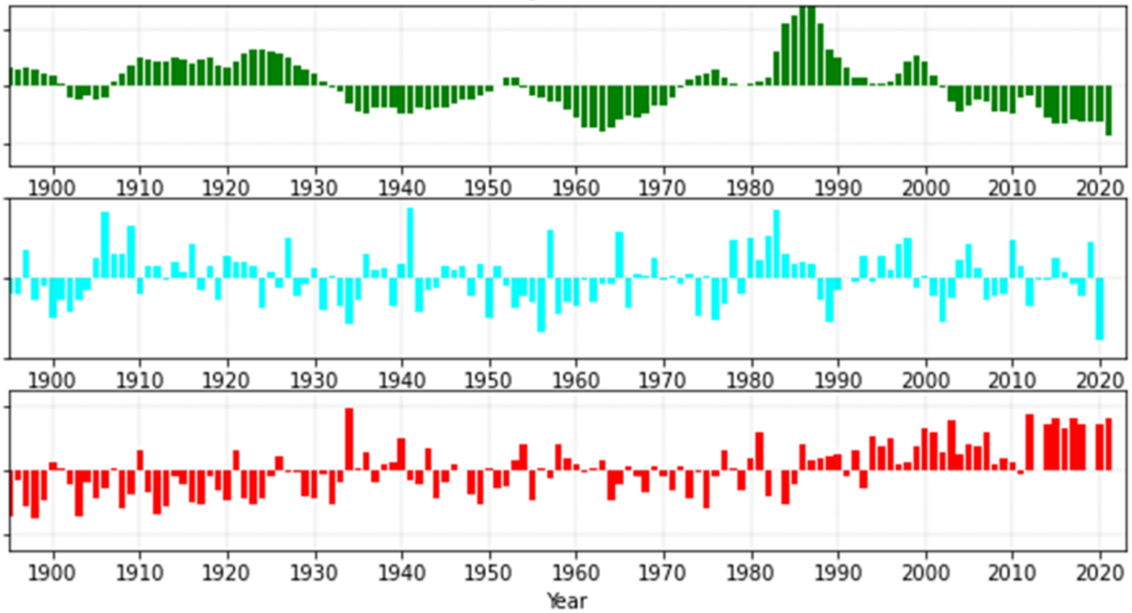

Figure 2.5. Top. Departure (m) of lake level from 127 year sample mean. Middle. Departure (cm) of Utah precipitation from sample mean. Bottom. Departure (°C) of Utah temperature from sample mean.

d. Transforming Data

The interpretation of a sample of data can be aided by transforming the data to a new variable. The simplest transformation is to remove the mean, in order to examine specifically the variability about a central value.

- $x_i' = x_i - \overline{x}$ = anomaly or departure from the mean

For example, the 127-year sample mean has been removed from the GSL levels in Fig. 2.5. Such a simple transformation can make a big difference as far as the interpretation of the data. In this instance, the anomalous period of the mid-1980's stands out. The recent period appears comparable to that during much of the middle part of the last century.

Strings of years with above and below normal precipitation in Utah are also more clearly evident in Fig. 2.5 than when the raw time series is examined, for example the wet years in the early 1980's. Removing the sample mean from the Utah temperature record reveals that the positive temperature anomalies since the 1990's are unprecedented. In addition, the 1934 temperature anomaly was quite unusual for that period.

Obviously, you can transform the data in a myriad of ways. What if we removed the current 1991-2020 "**climate normal**" for temperature instead (Fig. 2.6)? Then, the period prior to 1980 appears to be nearly always below normal for temperature and the years since 2010 are unusually high except recently in 2019.

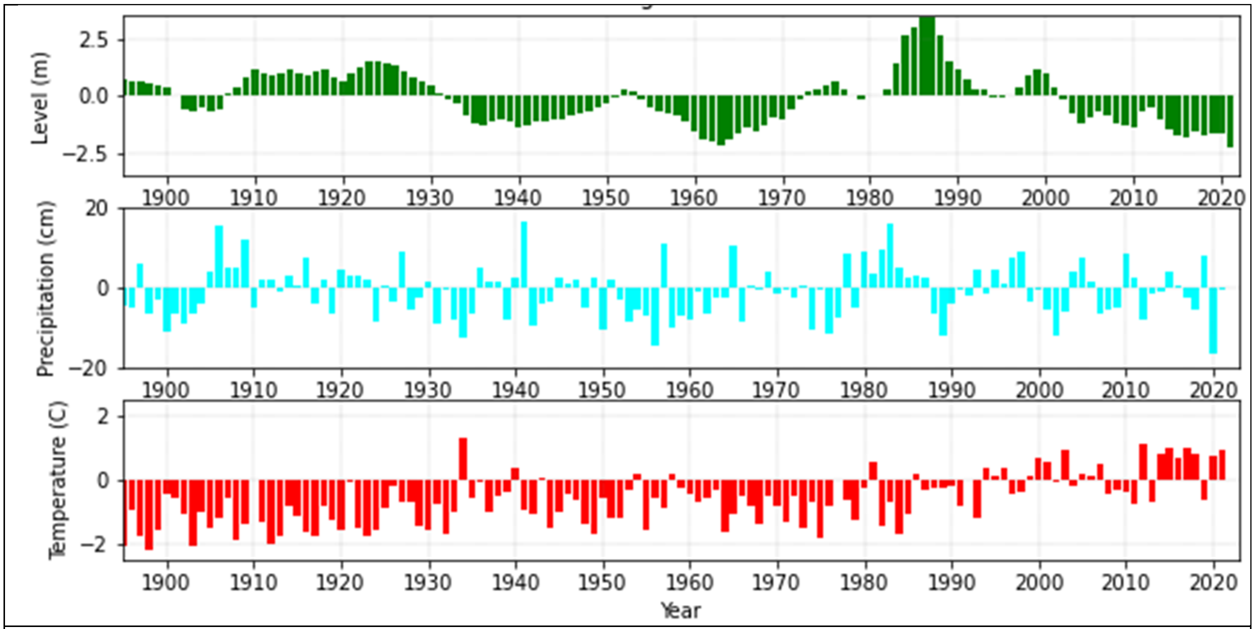

Figure 2.6. Anomalies relative to 1991-2020 climate normal.

In many environmental applications, quasi-periodic behavior due to solar forcing (diurnal or seasonal) can overwhelm the signal of interest. If you are interested in how much cooler each month in 2019 was compared to 2018, then the fact that January is always colder than July may not be relevant.

Let's examine the monthly Great Salt Lake level data since 1903 when the monthly values became available. You will need the data file gsl_monthly.csv. Figure 2.7 shows the **climatological monthly means** separately based on the 119 values available during each month. The lake level peaks on average in June (after the spring runoff period) and is lowest in the fall. Regular seasonal oscillations are detectable in the top panel of Fig. 2.8. Do we have 12*119 independent values, one for each month? No! Should we use the approach that the number of independent samples might be defined by the number of line segments required, do we now have 2*119 degrees of freedom (one for each year's rise and fall)? NO! We have 2 relevant time scales: the **annual cycle** as shown in Fig. 2.7 and the relatively small number of multi-year fluctuations already evident from the annually-averaged data. Those small number of multi-year fluctuations become apparent in the monthly data when the monthly means for each calendar month are removed as in the middle panel of Fig. 2.8.

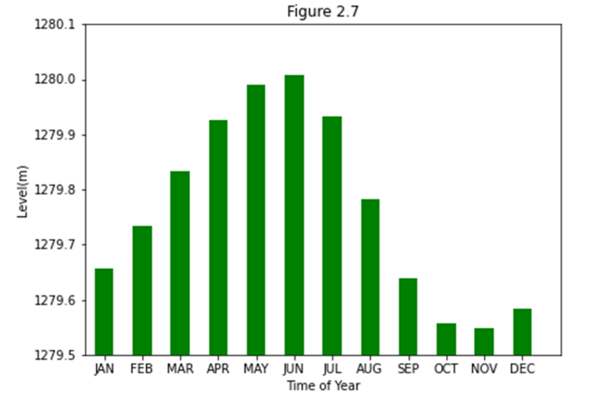

Figure 2.7. Lake level (m) averaged separately for each calendar month over the 118 years.

Since the variability in environmental data often differs during the year (e.g., weather systems moving across the midlatitudes are more frequent in winter than in summer), it is often appropriate when using data from all seasons to "normalize" or "standardize" the anomalies so that the variability in winter and summer receive similar weight. Hence, the value x_{ij} for year j and month i is subtracted from the mean for that month $\overline{x_i}$ and divided by the sample standard deviation s_{xi}, i.e.,

- $z_{ij} = \frac{x_{ij} - \overline{x_i}}{s_{xi}} = \frac{x'_{ij}}{s_{xi}}$

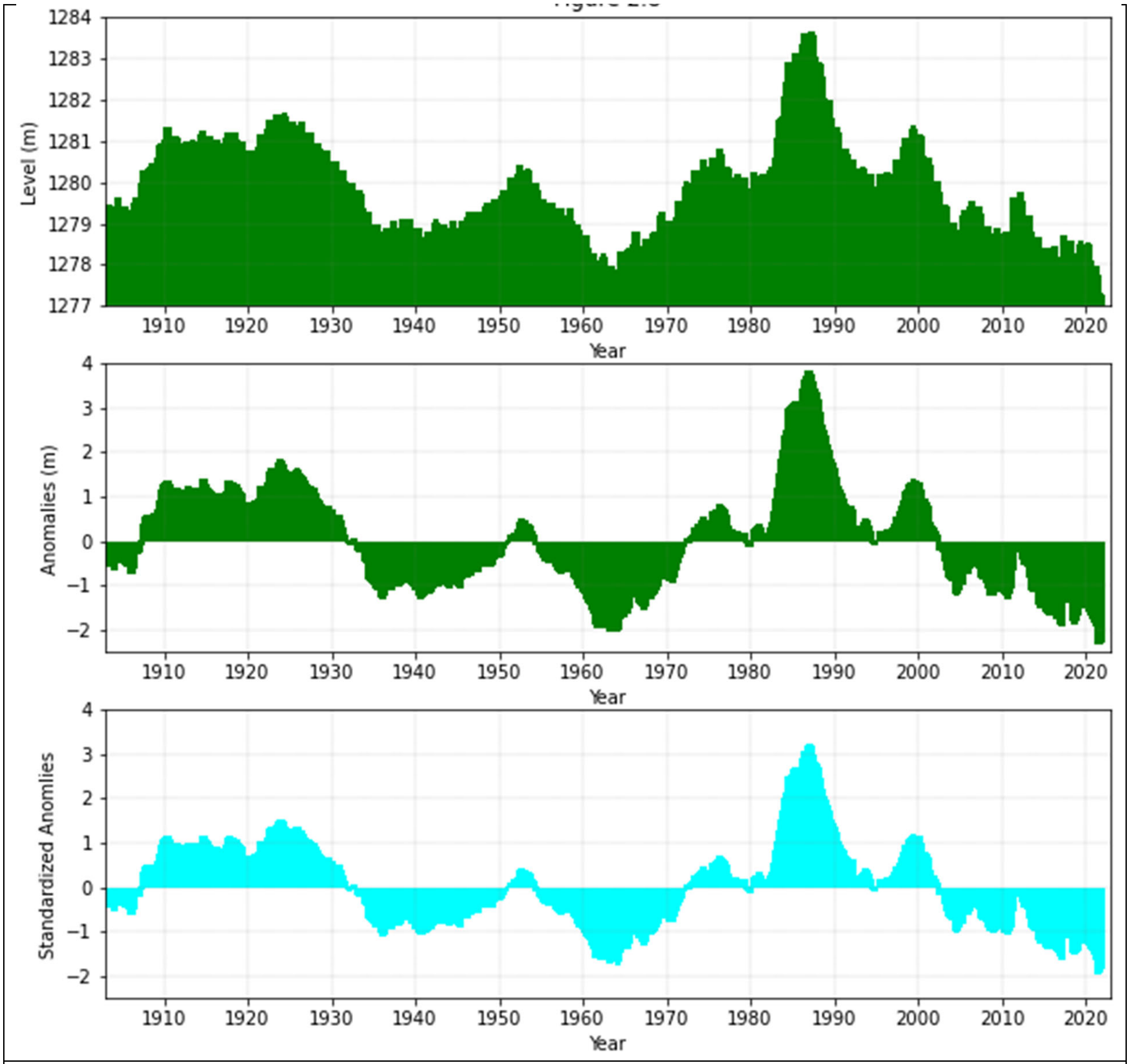

Figure 2.8. Top panel. Monthly lake level (m). Middle panel. Departures from monthly means (m). Bottom panel. Standardized anomalies (nondimensional).

In other words, a nondimensional time series is generated by creating '**standardized**' **anomalies**. As shown in the lower panel of Fig. 2.8, the lake level was three standard deviations above normal in 1986 and approached 2 standard deviations below normal in 1963. Even though we have a sample size of 1428 (119 years *12 monthly values), it is pretty clear from Fig. 2.8 that there remain ~20 independent samples in this time series, if we ignore some of the minor bumps and wiggles.

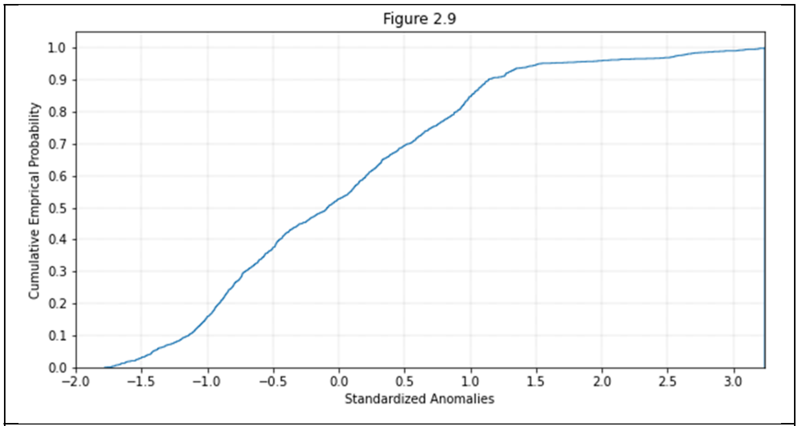

Figure 2.9. Cumulative frequency distribution for monthly values of lake level.

One of the advantages for appropriately transforming environmental data at higher sampling intervals (e.g., monthly vs. annually-averaged values) is that we can get an improved probability density distribution, as shown in Fig. 2.9. Compare this CDF to the one computed from the annual means. In terms of standardized anomalies, the median is 0, while the monthly lake level departs by more than ± 1 standard deviation around the mean cumulatively during 30% of the months. However, keep in mind that even though the probability distribution is smoother because of the larger sample size, the number of independent values remains low.

Variables such as wind speed and precipitation when examined from hour to hour or day to day tend to be strongly positively skewed. In other words, their distributions tend to be asymmetrical since no values are possible below 0, many values may be close to 0, and then occasional extreme values are possible. A simple transformation for wind speed or precipitation is to take the square root of the values, first, then remove the mean and standardize. For example, let

$y_{ij} = \sqrt{x_{ij}}$ and then $z_{ij} = \dfrac{y_{ij} - \overline{y_i}}{s_{yi}}$

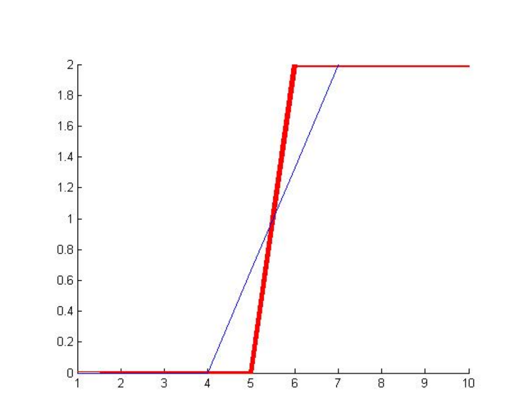

Figure 2.10. Example of original and median smoother (red line) and running mean smoother (blue line).

The annual Utah precipitation time series exhibits considerable year-to-year variability compared to the GSL level time series. A common transformation is to smooth a time series by redefining each element of the time series in terms of an aggregate of nearby values within a specified "filter window". The evaluation of megadroughts in Figure 1.1 used a 19 year window. Running means can shift peaks and smooth well-defined jumps in the data. The median smoother, where the median of values within a data window replace each element of the time series does a better job at maintaining sharp jumps in the data. Other simple weighting schemes will be shown later that can be used to filter out or retain specific components of the underlying time scales in the data (i.e., high pass, low pass, and band pass filters). A running mean filter is a low pass filter- the slower variations are "passed" and the higher frequency variations are removed.

Consider a portion of a time series on which a 3 point running mean and running median are applied as shown in Fig. 2.10. The original time series is the heavy line, i.e., the values are 0 from point 1-5 and then the data jumps to 2 at point 6. A three-point (or 5 point) median filter exactly matches the original data in this instance, which is sensible. However, a 3-point running mean (thin line) will smooth out the jump and lose the useful information that the sudden jump occurred at point 6. Instead you are left with the idea that the jump occurred slowly over 2 time intervals, not one.

As a general rule, it is better to aggregate the data within a small window and apply the smoother several times rather than aggregating the data within a large window and only applying the smoother once. Figure 2.11 shows the annual Utah precipitation anomalies after different combinations of window width (sample size) and interations. A yearly spike such as in 1941 is lost as "noise" while the anomalous string of wet years during the early 1980's begins to stand out as long as the sample size is not large. Is the window size of 11 really useful in this instance? Is it too smooth- what signal is left?

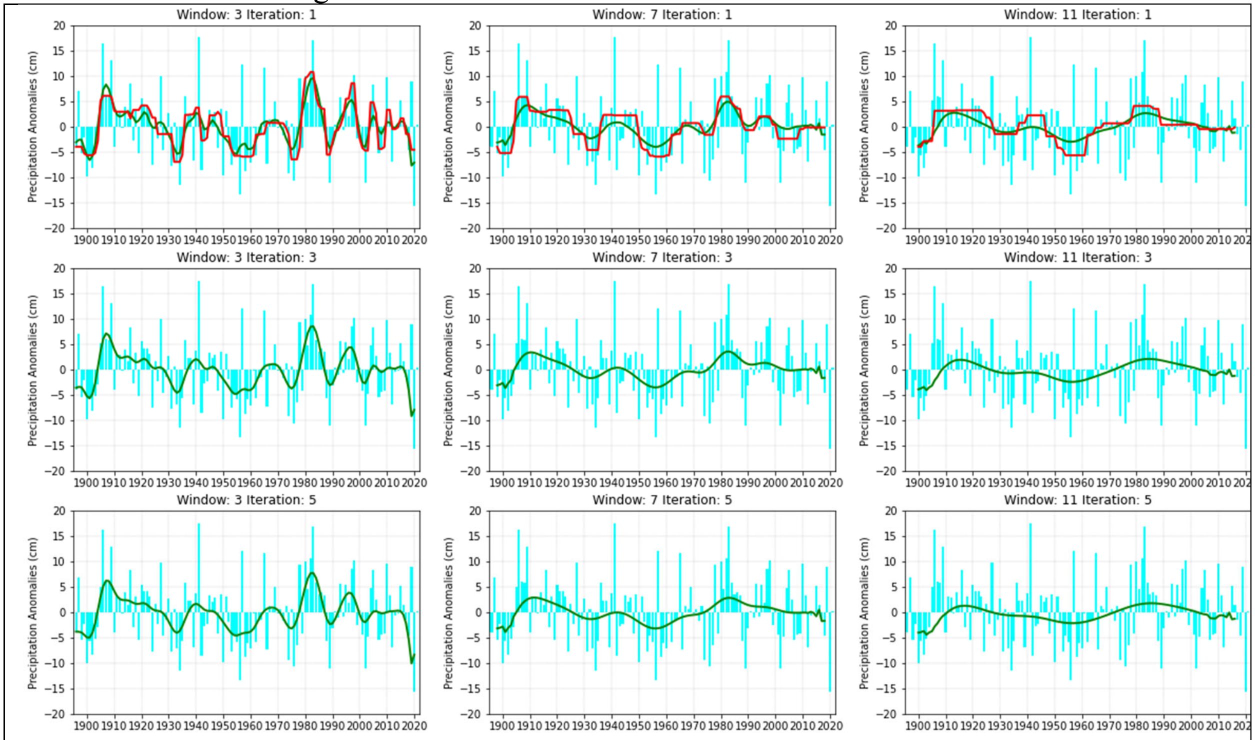

Figure 2.11. Utah annual precipitation anomalies (cyan bars) and the data smoothed using running mean smoother (green lines) using different windows (sample sizes; columns) and iterations (rows). The red lines in the top row indicate running median smoothers using different sample sizes (there would be no difference is iterated multiple times).

e. *Exploratory Univariate Analysis of Fluid Velocity*

Measures of central tendency and spread for two- and three-dimensional fluid motion can be analyzed univariately in terms of components of the flow, including vertical motion (w) and horizontal speed $\left|\vec{V}\right|$ and direction (θ) or horizontal Cartesian components, e.g., zonal u (east-west with u positive when fluid motion is from west to east) and meridional v (north-south with v

positive when fluid motion is from south to north). However, univariate metrics of fluid motion components can be misleading at times.

The horizontal fluid motion can be described in terms of the zonal and meridional wind as $\vec{V} = u\hat{\imath} + v\hat{\jmath}$ where the unit vectors define the positive directions (to the east and to the north) and the scalars (*u, v*) express the magnitude of the motion in that direction with a negative value indicating motion in the direction opposite to the unit vector. Alternatively, we can describe the horizontal wind velocity $\vec{V} = |\vec{V}|\hat{t}$ in terms of wind speed $V = |\vec{V}| = \sqrt{u^2 + v^2}$ where the unit normal $\hat{t}$ is tangent to the wind direction $\theta = 180 + \tan^{-1} u/v$ (θ is the direction from which the wind blows: north wind is 0; east wind is 90; south wind is 180; west wind is 270).

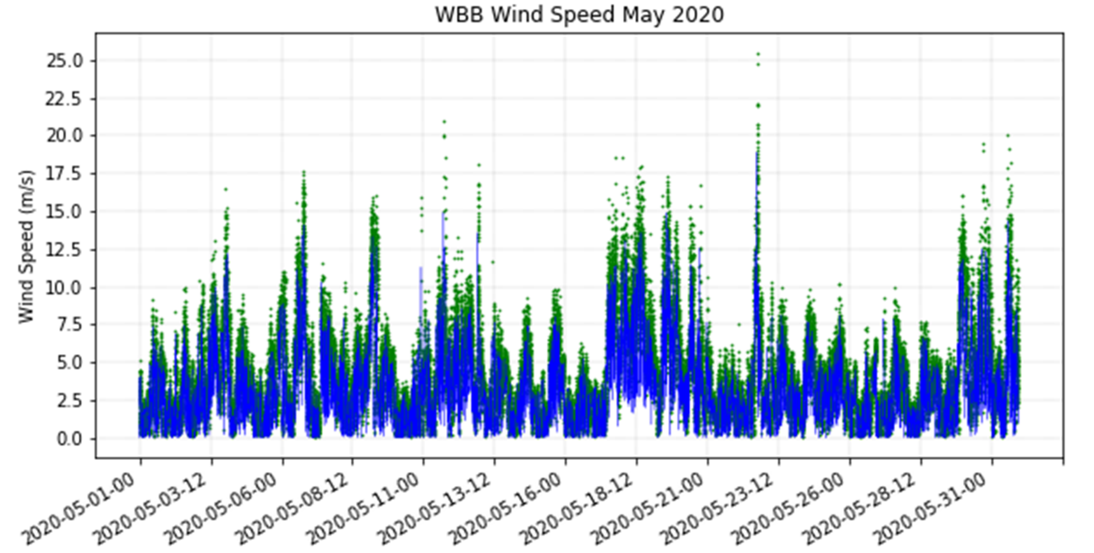

Figure 2.12. Wind speed (blue lines, m/s) and wind gust (green dots, m/s) during May 2020 on the tower atop the WBB. Times in MDT.

Consider the ~45,000 observations at 1 minute intervals of wind speed and wind gust in Fig. 2.12 from a wind sensor on the 10 m tower on top of the William Browning Building (WBB). Wind speeds are often light (less than 2.5 m/s, ~ 5 mph) but there were periods when weather systems moved through during which the winds were over 10 m/s. The wind gust metric reported from this station is the maximum or peak wind during the one-minute period. Those peak winds were as large as 25.5 m/s (~51 mph) at 6:02 PM May 25. Time series plots of wind direction are often difficult to interpret (1° and 359° are essentially the same but often plotted at opposite ends of the figure).

Reducing the dimensionality of those 45,000 observations could be done in many ways, but is most typically done in a way that loses much of the "weather": compute the average, minimum and maximum during each day. Histograms and cumulative frequency distributions can help retain some of the information lost by computing daily statistics. As shown in Fig. 2.13, the wind speeds at WBB are positively skewed with values > 6 m/s rare (only 10% of the time). Wind direction is multimodal (peaks around 45°-NE wind, 180°- south wind, and 315°- NW wind). To put this into a more physical context, winds on top of the Browning Building are strongly influenced by terrain flows, with nighttime winds from the northeast (away from the mountains) and daytime winds from the northwest (towards the mountains). Then, when weather systems are approaching, winds are frequently from the south. Note: in the original data there is a false peak at the zero angle, because there are some missing values interpreted as zero wind direction. That is handled in the python code and those missing values are removed.) The interpretation of the cumulative distribution of wind direction is not particularly useful, no more so than trying to interpret its time series.

Breaking the wind speed and direction into zonal and meridional components (Fig. 2.14) is occasionally useful, but not that useful in this situation. In this case, WBB winds were slightly skewed from the east (negative zonal wind resulting from the frequent downslope nocturnal winds from the east) while the meridional winds were positively skewed with winds more frequently from the south (positive meridional winds).

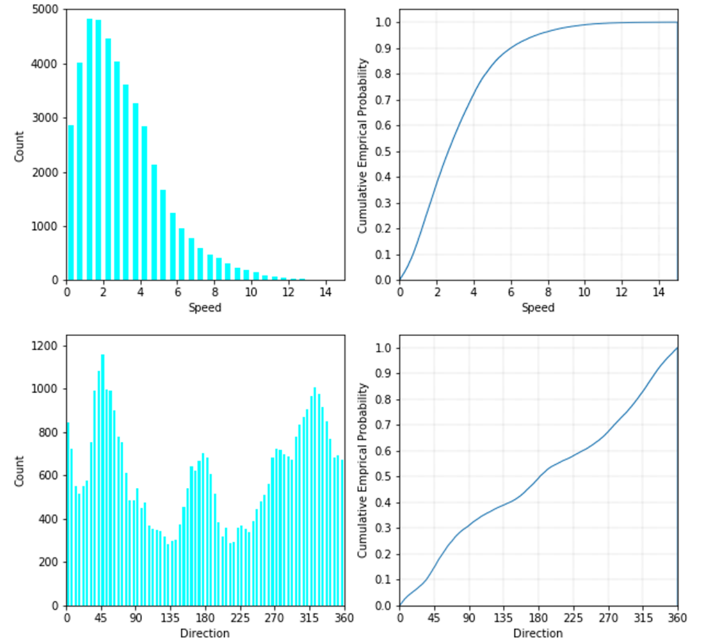

Figure 2.13. Histograms and cumulative frequency distributions of wind speed and direction.

As shown in Fig. 2.15, the wind rose is a special histogram that summarizes the counts from specified directions on a polar diagram. The python function breaks the counts into 1 m/s bins. Hence, winds were most frequent from the northeast, northwest, and south at the Browning Bldg. Strong winds were occasionally happening from all three of those directional ranges. The wind rose helps summarize the combined variability of wind speed and direction better than the individual histograms or cumulative distributions.

Particular care must be taken when computing an average of a vector. The mean wind speed $\bar{V} = \overline{|\vec{V}|} = \overline{\sqrt{u^2 + v^2}}$is not equal to the resultant wind speed defined as $\bar{V}_r = \sqrt{\bar{u}^2 + \bar{v}^2}$.

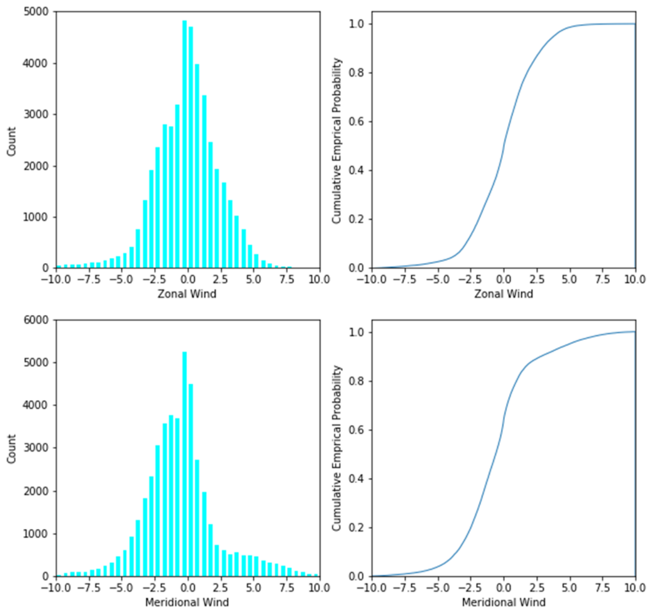

Figure 2.14. Histograms and cumulative frequency distributions of zonal and meridional wind components.

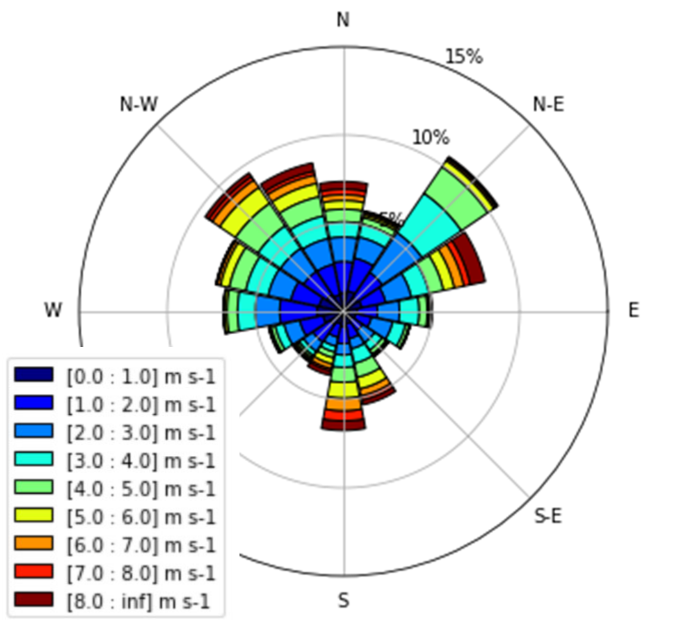

Figure 2.15. Wind rose of WBB winds during May 2020. Percentage of values in 22.5° bins in 1 m/s intervals.

Fig. 2.16 compares the mean and resultant wind speeds as a function of time of day. The strongest winds were found in the afternoon (15-16 MDT) and evening while the weakest winds tend to occur early in the morning. The resultant wind speed is clearly less than the mean wind speed each hour.

Taking a simple average of wind direction should also be avoided, since, for example, the average wind direction for observations of 359° and 1° should be 0° not 180°. The average wind direction loses much of its meaning when wind speed is not considered at the same time. A common procedure is to compute the resultant mean wind direction as $\theta_r = 180 + \tan^{-1} \overline{u} / \overline{v}$.

Then, it is possible to plot the resultant vector $\vec{V_r} = \left|\vec{V_r}\right| \hat{t}_r$where the tangential unit vector is defined from the resultant mean wind direction.

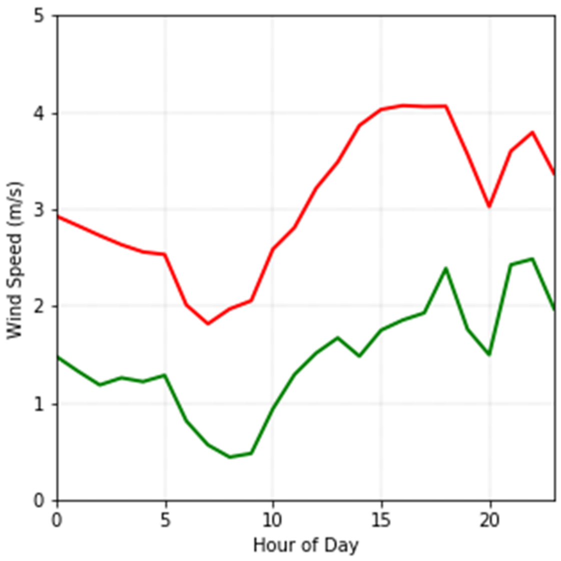

Figure 2.16. Comparison of hourly mean wind speed (red line) and resultant wind speed (green line) for each hour of the day at WBB during May 2020.

3 a. Probability

a. Definitions

We are inundated with probabilities in environmental fields as well as society. The chance of rain is 50%- what does that mean? The chance of lung cancer in bald males who smoke is XX%. Probabilities should be defined carefully. We begin with some definitions.

- Event- set or class or group of possible uncertain outcomes. Rain/no rain. Temperature greater than 50ºF, etc.
- Elementary event- cannot be decomposed into other events
- Compound event- decomposable into 2 or more elementary events or other compound events
- Null event- that which cannot occur

Example: roll 6 sided die. (1) elementary event- 1 spot comes up; (2) compound event- odd number of spots comes up (1, 3, or 5); (3) null event- getting a 7 on a 6 sided die.

Will precipitation occur tomorrow? That is an elementary event if the only other choice is no precipitation. However, a compound event would be: will precipitation greater than 0.1 inch occur (it could rain more or could rain less or not at all) or will it snow or rain or both?

- S- Sample or event space. Set of all possible elementary events or the largest possible compound event
- Mutually exclusive- two events that cannot occur at the same time
- Mutually exclusive and collectively exhaustive events (MECE)- no more than 1 event can occur and at least one event will occur

b. Venn Diagrams

Venn diagrams are a convenient way to display the sample space and make sense of the event outcomes that are possible. The NCDC storm event climatology (http://www.ncdc.noaa.gov/stormevents/) is a rich resource for examining weather events. From the reports for Salt Lake County from the NCDC Storm Event climatology, the number of cases were tabulated of winter and summer (convective) storms and those storms with property damage greater than $5000 during the thirteen year period 1993-2005. Now, some assumptions were made along the way as far as how to count events- some winter storms events may have been multiple day events, for example, and lightning occurrences were associated with convective storms. Some iffy cases where ignored that could have been a convective winter storm. Property damage has occurred from "other" storms and obviously the results might have been different if another $ damage threshold was used. In any event, there were a total of 142 winter storms and 83 summer storms as defined. 79 winter storms had damage in excess of $5000 while 25 summer storms had damage of similar amount. Given the nearly 5000 days during the 13 year record, these major weather events as defined by NCDC are not very common in Salt Lake County. The Venn diagram helps to highlight that winter storms are associated with property damage more frequently than summer storms in Salt Lake County and, as defined here, winter and summer storms are obviously mutually exclusive.

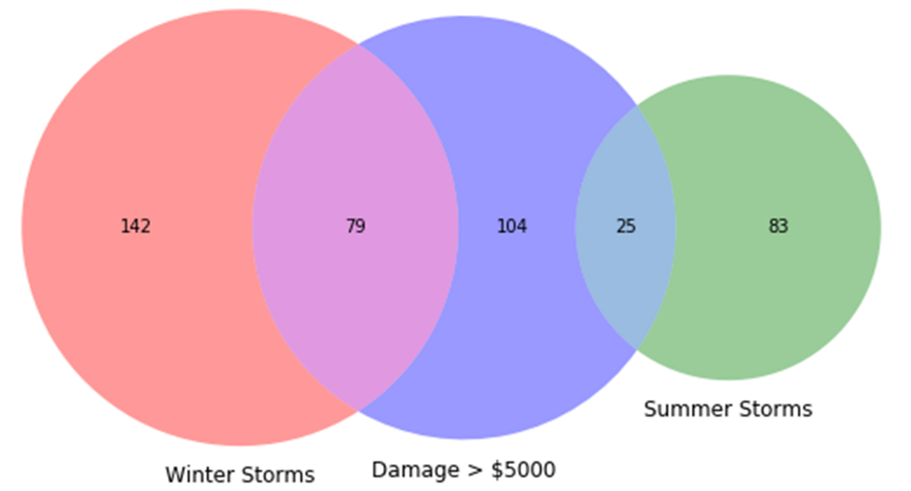

Figure 3.1. Venn diagram of major storms in Salt Lake County during 1993-2005.

Venn diagrams are useful for categorizing events that fall into clear categories and they don't need to be done in terms of circles. Consider Fig 3.2 that shows the possible MECE for a seasonal forecast of above/below normal temperature and precipitation for a specific location. All four possibilities are shown and the probability of each event will depend on the situation and location.

Seasonal Forecast Events

Temperature below Precipitation below	Temperature above Precipitation below
Temperature below Precipitation above	Temperature above Precipitation above

Figure 3.2. MECE possibilities for seasonal forecasts of temperature and precpitation anomalies for a specific location.

c. *Probability Concepts*

The following are pretty obvious, but when you get mathematicians involved, they have to have "axioms", lemmas, etc.

- probability of any event is nonnegative. In English: an event has to happen or else it is not an event
- probability of the compound event S is 1 or 100%. The probability that all events will happen is 1.
- probability that one or the other of two mutually exclusive events is the sum of their individual probabilities.

Definitions:

- Let E- event
- Pr{E}- probability of Event E; $0 \le \Pr\{E\} \le 1$
- Pr{E}=0 event does not occur
- Pr{E}=1 absolutely sure that event will occur

There are two approaches to probabilities: the frequency view and the Bayesian view. Which approach is used depends on the type of problem being investigated.

Frequency view- probability of an event is its relative frequency after many, many trials

- a- number of occurrences of E
- n- number of opportunities for E to take place
- a/n – relative frequency of event E occurring
- $Pr\{E\} \rightarrow a/n$ as $n \rightarrow \infty$
- Or a = outcomes = n Pr{E}

Examples: role a die. We expect the 6 spot to come up 1/6 times or 1 time every 6 opportunities. If we role the die 100 times, we expect the 6 to come up 16-17 times. However, what we expect and what actually happens are clearly different things, that's where chance/randomness comes into play.

Bayesian view- probability represents the degree of belief or quantifiable judgement of a particular individual about an outcome of an uncertain event

- this approach recognizes that some events occur so rarely that there is no long-term probability estimate that are relevant
- Bookies make odds all the time based on their evaluation of the odds of winning for a particular team- it is not based on a large sample
- Two individuals can have different probabilities for same outcome

More concepts

- If event $\{E_2\}$ occurs whenever $\{E_1\}$ occurs, then $\{E_1\}$ is a subset of $\{E_2\}$
- Example: $\{E_1\}$- temperature below freezing; $\{E_2\}$- temperature below 50F, then $Pr\{E_1\} \leq Pr\{E_2\}$
- The complement of {E} is that event $\{E\}^c$ that does not occur
- $Pr\{E\}^c = 1 - Pr\{E\}$

What is the probability that $\{E_1\}$ and $\{E_2\}$ occur, that is, the intersection between the two events?

- $Pr\{E_1 \cap E_2\}$ = joint probability that $\{E_1\}$ and $\{E_2\}$ will occur *(3.c.1)*
- $Pr\{E_1 \cap E_2\} = 0$ if $\{E_1\}$and $\{E_2\}$are mutually exclusive
    - Example: if $\{E_1\}$is the occurrence of temperature below freezing and $\{E_2\}$ is the occurrence of temperature above 50°F, then their joint probability is 0.

Let's return to the Venn diagram of the weather events in Salt Lake County. In the way that the sample was created, the winter storms and convective storms are mutually exclusive, so there is no overlap between those two events. Assuming, that winter storms occur only during the winter half of the year and that convective storms occur only in the summer half (not great assumptions!), then the number of opportunities is order 180 days x 13 years= 2340 opportunities. Also, remember that some of the winter storms could be multiple day events, so there is some uncertainty and error in the results.

- $\{E_1\}$- occurrence of winter storms = 142
- $Pr\{E_1\}$= 142/2340 = .061
- $Pr\{E_1\}^c$ = .939

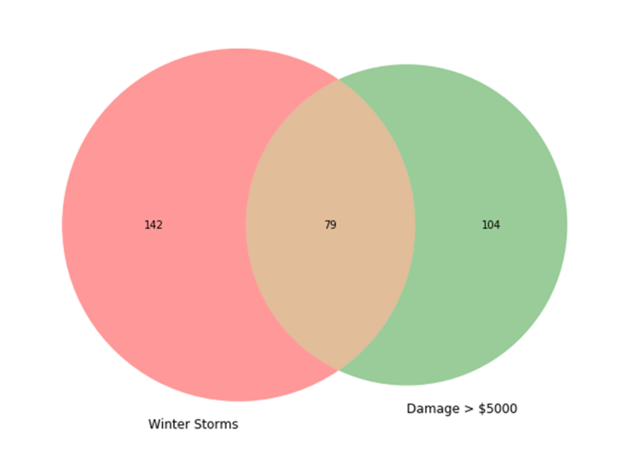

Figure 3.3. The overlap (Pr{$E_1 \cap E_2$}) has to be subtracted from the sum of the probabilities of each event occurring in order to determine if either occur (Pr{E_1 U E_2}).

- {E_2}- occurrence of summer convective storms = 83
- Pr{E_2}= 83/2340 = .035
- Pr{E_2}^{c} = .965
- {E_3}- occurrence of property damage = 104
- Pr{E_3}= 104/2340 = .044
- Pr{E_3}^{c} = .956
- Pr{$E_1 \cap E_2$} = 0
- Pr{$E_1 \cap E_3$} = 79/2340 = .034
- Pr{$E_2 \cap E_3$} = 25/2340 = .011

What is the probability that {E_1} OR {E_2} will occur? That is, one the other, or both will occur. This is referred to as the "union" of the two events.

- Pr{$E_1 \cup E_2$} = Pr{E_1} + Pr{E_2} - Pr{$E_1 \cap E_2$} *(3.c.2)*
- As can be seen visually from the Venn diagram to the right, the joint probability is counted twice when the individual probabilities are summed, so it is subtracted once.

It is worthwhile to see how that is true algebraically as well. Add up each probability separately:

Pr{$E_1 \cup E_2$} = Pr{E_1} - Pr{$E_1 \cap E_2$} + Pr{$E_1 \cap E_2$} + Pr{E_2} - Pr{$E_1 \cap E_2$} or
Pr{$E_1 \cup E_2$} = Pr{E_1} + Pr{E_2} - Pr{$E_1 \cap E_2$}

- If {E_1} and {E_2} are mutually exclusive, then Pr{$E_1 \cup E_2$} = Pr{E_1} + Pr{E_2}

A couple more identities that you should be able to visualize from a Venn diagram:

- Pr{$(E_1 \cup E_2)^c$} = Pr{ $E_1^c \cap E_2{}^c$}
  - that is the area outside of both circles
- Pr{$(E_1 \cap E_2)^c$} = Pr{ $E_1^c \cup E_2{}^c$}
  - this one is a little harder to visualize; it is everything outside of the intersection of the two events

Returning to the Salt Lake County storm data:

- Pr{$E_1 \cup E_2$} = Pr{winter storms} + Pr{convective storms} - Pr{winter storms ∩ convective storms} = .096 since the two events are mutually exclusive the last term is 0
- Pr{$E_1 \cup E_3$} = .061 + .044 - .034 = .071 = Pr{winter storms or property damage or both}
- Pr{$E_2 \cup E_3$} = .035 + .044 - .011 = .068 = Pr{summer storms or property damage or both}

d. Conditional Probability

The storm data example for Salt Lake County indicates that some winter storms lead to expensive damage. So, given that a winter storm has occurred, what is the probability that damage has occurred? We are now limiting our sample to a smaller number of events, only the 142 winter storms. So, the probability is now the 79 damaging winter storms divided by the 142 total winter storms or 56%.

- Conditional probability: probability that $\{E_2\}$ will occur given that $\{E_1\}$ has occurred
- $\Pr\{E_2 \mid E_1\} = \Pr\{E_1 \cap E_2\} / \Pr\{E_1\}$ *(3d.1)*

$\{E_1\}$ is called the conditioning event; if it doesn't happen, then we know nothing about the probability that $\{E_2\}$ will happen

Alternatively, we can write:

- $\Pr\{E_1 \cap E_2\} = \Pr\{E_2 \mid E_1\} \times \Pr\{E_1\} = \Pr\{E_1 \mid E_2\} \times \Pr\{E_2\}$ *(3d.2)*

Whether we condition from the first or second event to determine the intersection of the two events is our choice and simply depends on the available data.

If two events are completely independent, such that the occurrence of nonoccurrence of one event does not affect the probability of the other, then

- $\Pr\{E_2 \mid E_1\} = \Pr\{E_2\}$ and $\Pr\{E_1 \mid E_2\} = \Pr\{E_1\}$

Then,

- $\Pr\{E_1 \cap E_2\} = \Pr\{E_1\} \times \Pr\{E_2\}$ for independent events

If we have a fair coin, then the Pr{head} = .5. The second coin toss does not depend on the first, so Pr{10 heads in a row} = $.5^{10}$

e. Persistence

Persistence is the existence of statistical dependence over time (or space), i.e., that once a phenomenon begins it does not necessarily end before the next observation time or, that the observations at one location are related to the observations at a nearby one. Observations from environmental fields should not be considered to be independent of one another unless care is taken to choose a sample taking into account spatial and temporal dependence.

Consider the fog climatology shown in Fig. 3.4 that was created for the 2002 Olympics by Jonathan Slemmer. We wanted to provide the Olympic forecasters with some information on the likelihood of persistent heavy fog at the airport. If it happened during the Olympics, then there would have been a bunch of negative consequences with flight delays, etc. (fortunately, it didn't happen during that period). Dense fog doesn't happen very often. The sample here is large: 31 years x 365 days= 11315 days. Dense fog happened over a 2 hour period (Jonathan labels this 1 h of consecutive fog) on only 202 days. So Pr{ 1 hour of consecutive fog} = Pr{1}= 202/11315

= 1.8% of days. If we considered the number of hours in a day as well, then the probability that it would happen in a specific hour would be correspondingly less. Pr{2} = 109/11315 = .96% of days, etc.

The probability that 3 hours in a row of fog (2 consecutive hours) is going to happen is pretty unlikely = .96% for any given day. However, if 1 hour of consecutive fog has already happened, then the odds of it continuing are obviously much higher.

- Pr{2|1} = 109/202 = 53.9%, Pr{3|1} = 70/202 = 35%, Pr{10|5} = 41.6%, etc.

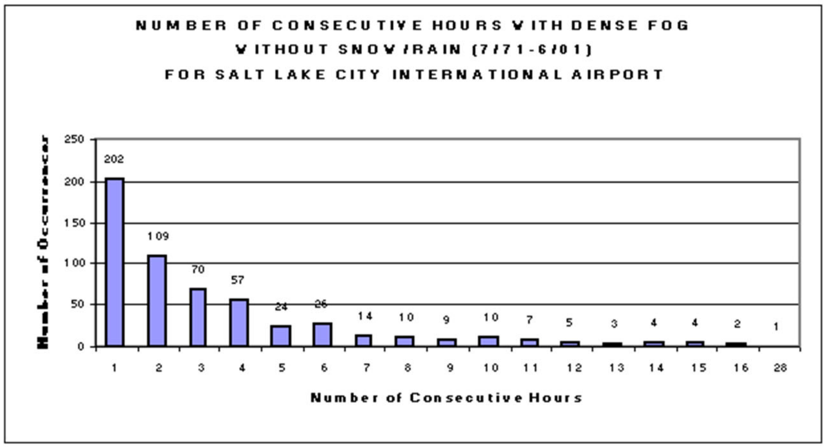

Figure 3.4. Fog climatology at the Salt Lake airport.

Persistence is a good statistical forecasting baseline. When my family asks what the weather is going to be like tomorrow, without any other information available, I'm going to say whatever it is today. A forecaster adds value when the conditions are present that lead to change and those conditions are correctly recognized as such. While it may be useful to tell an airport operator that the current dense fog has a high likelihood of continuing, more value will be added if the forecaster has information available from which to diagnose when the fog is going to break up.

Later, we will examine ways to estimate the probability of rare events. For example, it is not particularly useful to develop probabilities on the occurrence of dense fog for over 20 consecutive hours based on the 1 event that has happened in our sample. Similarly, we can't wait around for a 100 years to estimate the occurrence of a once in a hundred year flood.

f. Forecast Verification

We'll touch on verification of forecasts from several angles. The differences are introduced here between a "measures oriented verification approach" (typically using well-established and often over-used performance metrics such as hit rate or threat score) vs. a "distributions-oriented approach" (where the empirical joint distributions of forecasts and verifying observations are generated).

Let's start with the simple approach that something happens or it doesn't and we forecast it to happen or not. We then count the number of cases for the four possibilities. First, the marginal totals are important- they are how often something is observed or forecast (or not observed/not forecast). How often do we get the "right" forecasts and how often do we forecast them and they don't happen?

PC = percent correct = $\frac{a+d}{n}$

FAR = false alarm ratio = $\frac{b}{a+b}$

		Observed	Observed	**Forecast marginal totals**
		Yes	No	
Forecast	Yes	a	b	**a+b**
Forecast	No	c	d	**c+d**
	Observed marginal totals	**a+c**	**b+d**	**n=a+b+c+d sample size**

We want the percent correct to be close to 1 and the false alarm ratio to be close to 0. But, what happens when we are verifying categorically forecasts of large precipitation amounts (say over an inch) at Salt Lake City? That doesn't happen very often, but the percent correct may be large because we may be very successful at forecasting when the precipitation is less than an inch (d). Then, to focus on the situations when it either was observed or forecast, the threat score (TS) or critical success index (CSI) is used:

TS = CSI = $\frac{a}{a+b+c}$

The threat score measures the number of correct "yes" forecasts relative to the total number of occasions on which the forecast was forecast or observed. Another metric commonly used is the probability of detection (POD) or hit rate (HR), which identifies how frequently an event is forecast relative to when it is observed:

POD = HR = $\frac{a}{a+c}$

Note when we are using one of the marginal totals in the denominator, we're computing a conditional probability. We can express the hit rate as: given that an event occurs, how often is it correctly forecast? The FAR is: given that an event is forecast, how often did it not happen?

Now let's look at an example using Matt Lammer's research on verifying forecasts made in support of prescribed burn and wildfire operations: http://meso1.chpc.utah.edu/jfsp/

On January 30, 2015, there were 77 forecasts issued in support of prescribed burns nationwide. How often did the forecasters anticipate high wind speeds (≥ 5m/s) later that afternoon to take place relative to what was observed that day?

		Observed	Observed	**Forecast Marginal totals**
		≥ 5m/s	<5 m/s	
Forecast	≥ 5m/s	11	6	**17**
Forecast	<5 m/s	16	44	**60**
	Observed Marginal totals	**27**	**50**	**77**

So, the PC= 71.4%; FAR= 35.3%; TS= 33.3%; and POD = 40.7% . How did they do? Clearly, the percent correct is high because they forecast correctly a lot of cases when the winds were

light. The false alarms are not too bad, 6 of the 17 forecasts of high wind. But, the threat score is low because they missed a lot of cases when the winds were observed to be stronger than they were expecting.
Do these forecasts have skill? Statistical skill refers to the relative accuracy of a set of forecasts with respect to reference forecasts (random, persistent, or climatological, for example). The probability of a correct yes forecast by chance (meaning that the observations and forecasts are independent) is just the product of the marginal probabilities of the observations and forecasts:
Random correct yes forecast by chance = $\frac{(a+b)}{n}\frac{(a+c)}{n}$
Random correct no forecast by chance = $\frac{(b+d)}{n}\frac{(c+d)}{n}$

In our case, the odds of having a randomly correct yes forecast is low (7.7%) but the odds of having a randomly correct no forecast is pretty high (50.1%) since it is both observed and forecasted to not be windy frequently.

The most generic of skill scores is of the form: $SS = \frac{(correct\ forecasts - random\ correct\ forecasts)}{(total\ forecasts - random\ correct\ forecasts)}$
The Heidtke Skill Score is of this form and can be computed after some substitutions from the contingency table values as: $HSS = \frac{2(ad-bc)}{(a+c)(b+d)+(a+b)(b+d)}$

In our case, HSS= 31.4%, which is not particularly high and reflects that low wind speed forecasts don't require a lot of skill.

The measures-oriented metrics defined above are ok, but much information is lost by looking only at 2x2 contingency tables. Let's broaden the scope a bit and assume that an accurate forecast is one when the forecast wind speed is within 2 m/s of the observed forecast. So, most frequently, the forecast errors are up to one m/s weaker than those observed. And, as we determined from the earlier metrics, there is a greater tendency to forecast the wind speeds to be lower than those observed on this particular day. However, we don't know from Fig. 3.4 whether the forecasters do a better job over some ranges of observed wind speeds than others. We can expand the contingency table concept to create a "distributions-oriented" approach to verification as shown in the following table. I've now arbitrarily decided an accurate forecast is within ±2 m/s of that observed and then keep track as well of the sign of the errors that exceed that limit. I've broken up the observed wind speeds into 3 categories as well. Now it is clearer that the forecasters tend to underforecast higher wind speeds and don't ever overforecast high wind speeds, only more moderate ones.

This is a MECE data set for this particular sample of forecasts issued on this single day. If we were to divide the counts in the interior bins by the sample size (77), then those interior bins would be joint probabilities, e.g., 26% of the forecasts were within 2 m/s when the wind speeds were between 3 and 6 m/s (20/77). A lot more information can be gleaned by considering the conditional probabilities as defined by 3c.1 and 3c.2. For example, given that the observed wind speed is greater than 6 m/s ($\Pr\{E_1\}$ = 18/77= 23.4%), the probability that the forecasters predict the winds to be too light $\Pr\{E_2 | E_1\}$ is:

$\Pr\{E_2 | E_1\} = \Pr\{E_1 \cap E_2\} / \Pr\{E_1\}$ = ((11/77)/(18/77)) = 64.7%

And, here’s where the interpretation of conditional probabilities can get out of hand. On this particular day, given that the error is greater than +2 m/s, then the probability that the wind is in the 3-6 m/s category is 100%!!

$Pr\{E_2 | E_1\} = Pr\{E_1 \cap E_2\} / Pr\{E_1\} = ((7/77)/(7/77)) = 100\%$

Table 3.1. Distribution-oriented verification of wind forecasts

	E_1	Observed	Observed	Observed	**Error Marginal totals**
E_2		≤3 m/s	3-6 m/s	≥6 m/s	
Error	≤ -2 m/s	0	10	11	**21**
Error	± 2 m/s	22	20	7	**49**
Error	> 2 m/s	0	7	0	**7**
	Observed Marginal totals	**22**	**37**	**18**	**77**

Hooray, we can say all forecasters tend to overforecast high winds when the winds are between 3-6 m/s (no- we can’t).

Now, imagine only one in ten thousand people will get a particular disease- $Pr\{E_1\}$. But you hear on the news that 50% of the people that come down with the disease ate jello that day- $Pr\{E_2$- ate jello $| E_1\}$. Should you stop eating jello to avoid catching the disease? $Pr\{E_1 \cap E_2\} = Pr\{E_2 | E_1\} \times Pr\{E_1\} = .50 * .0001 = .005\%$ Don’t focus on that eating jello seems to cause an alarming increase in risk; the more important issue is the low risk factor for this particular disease under any circumstance.

g. Summary

Probabilities are at the heart of modern weather forecasting as well as many other environmental applications. While many applications and users will continue to expect to hear on the radio what the temperature will be at 4 PM tomorrow, the underlying information from which a forecaster will base that specific number will likely be probabilistic information. For example, forecasters implicitly use conditional probabilities as part of the forecast preparation. Given the approaching front, and given that a specific model has a known bias in temperature in the prefrontal environment, they expect the temperature to be higher/lower than what would normally take place.

3b. Theoretical Distributions and Hypothesis Testing

a. Parametric and Empirical Probability Distributions

The empirical histograms and cumulative density distributions discussed in Chapter 2 have many applications but they are determined from a sample of the population. Parametric probability distributions are a theoretical construct using mathematical relationships to define populations with known properties. One or two parameters combined with the assumption that the population is composed of random events may be enough to define the occurrence of possible outcomes of an environmental phenomenon. By comparing parametric and empirical probability distributions, we can deduce additional information about the population from which a sample is taken. The advantages of applying parametric distributions include:

- **compactness**- we may be able to describe a critical aspect of a large data set in terms of a few parameters
- **smoothing and interpolation**- our data set may have gaps that can be filled using a theoretical distribution
- **extrapolation**- because environmental events of interest may occur rarely, our sample may not contain extreme events that could be estimated theoretically by extending what we know about less extreme events

But keep in mind that while parametric distributions have advantages, they also can instill a level of confidence about your understanding of a phenomenon out of proportion to what really can be known. For example, as part of the Chapter 2 assignment, you estimated Utah's average temperature in 2050 from extrapolating a linear trend. That may or may not be a good idea.

Roman letters (e.g., s- sample standard deviation) are used to define sample statistics while Greek letters (e.g., σ- population standard deviation) are used to define the population statistics. Since parametric probability distributions are a theoretical construct that hopefully describes the population, the parameters used to define them are generally given by Greek letters.

Many environmental phenomena are discrete events: it either rains at a particular location or not; a tornado touches down or not; an earthquake happens in a location/time or it doesn't. There are a large number of parametric distributions (binomial, Poisson, etc.) appropriate for examining a data set of discrete events. Because of the limited time available in this course, we are not going to discuss discrete parametric distributions. On the other hand, most environmental variables of interest can be defined as being continuous: whether it rains or not is part of a continuum of how much it rains; we can classify temperature above or below a threshold as a discrete event but temperature varies continuously over a wide range of values; earthquake intensity is defined continuously on the Richter scale. There are a suite of parametric distributions (Gaussian, lognormal, gamma, Weibull, etc.) that are relevant to continuous distributions.

It is important to recognize the steps involved in using parametric distributions:

- generate an empirical CDF
- determine a good match between the empirical CDF and a particular parametric distribution

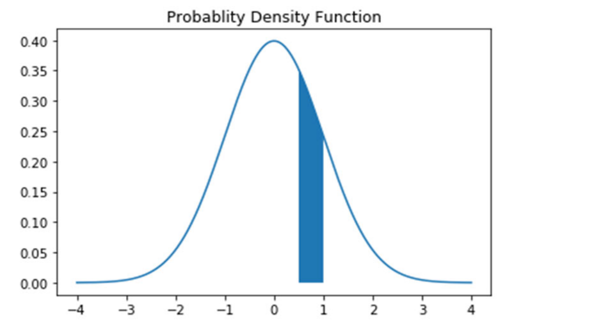

Fig. 3.5. Probability density function.

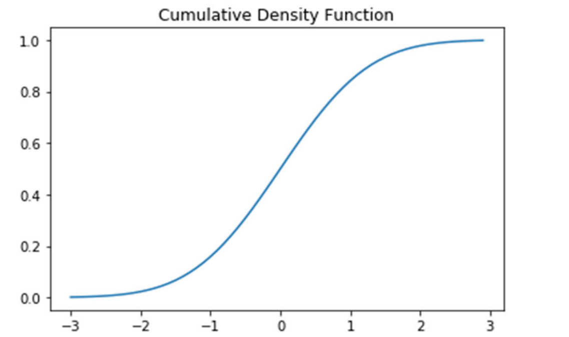

Fig. 3.6. Cumulative density function

- use the parameters from that parametric distribution to estimate the probabilities of values above or below a threshold, likelihood of extreme events, etc.

We begin by defining the probability density function (PDF) for a random continuous variable x as f(x), which is the theoretical analog of the histograms in Chapter 2. The sum of f(x) over all possible values of x is $\int_{-\infty}^{\infty} f(x)dx = 1$. As with the interpretation of integrals in general, think of the product $f(x)dx$ as the incremental contribution to the total probability. The shaded area shown in Fig. 3.5 represents $\int_{.5}^{1} f(x)dx$ and represents 15% of all the possible values. The cumulative distribution function (CDF) is the total probability below a threshold, hence, the total area to the left of a particular value: $F(X) = \Pr\{x \le X\} = \int_{-\infty}^{X} f(x)dx$. For example, for the CDF in Fig. 3.6, the cumulative probability of negative values is 50%. Also, it is useful to define X(F) as the value of the variable corresponding to a particular cumulative probability, e.g., from the figure X(75%)=0.66.

The function that defines all possible values of X(F) is referred to as the quantile function. The expected value, E, of a random variable or function *of a random variable* is the probability-weighted average of that variable or function.

- $E[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx$

Consider this intuitively as weighting the values of g(x) by the probability of each value of x. A reminder of a few integral properties:

- for a constant c, $E[c] = c$ since the sum of f(x) over all values of x is simply 1
- for g(x)=x, $E[x] = \int_{-\infty}^{\infty} xf(x)dx = \mu$: μ is the mean of the distribution whose PDF is f(x)
- $E[cg(x)] = c\int_{-\infty}^{\infty} g(x)f(x)dx$

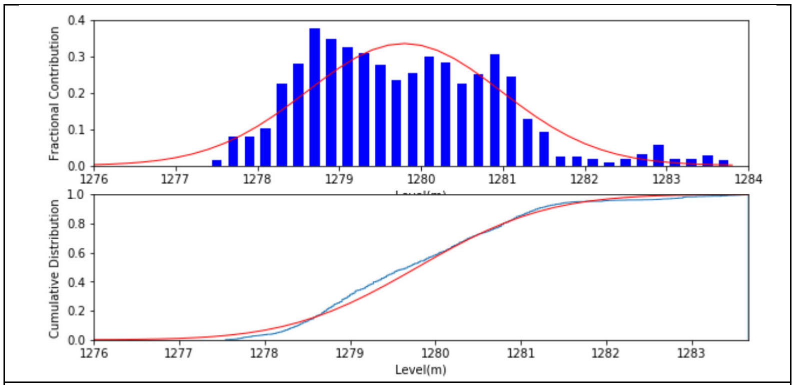

Figure 3.7. Gaussian fit to the monthly level of the Great Salt Lake in terms of its histogram (top) and cumulative distribution (bottom).

- The contribution to the total variance from a particular value of x is $g(x) = (x - E(x))^2$. So, the total variance is

$$Var[x] = E[g(x)] = \int_{-\infty}^{\infty} (x - E(x))^2 f(x)dx = \int_{-\infty}^{\infty} (x^2 f(x)dx - 2xE(x)f(x)dx + E(x)^2 f(x))dx$$

$$= E(x^2) - (E(x))^2 = E(x^2) - \mu^2$$

We'll use the above relationships for several different continuous parametric distributions.

b. Gaussian Parametric Distribution

Each parametric distribution that you are likely to use has a rich tradition in statistics, none more so than the Gaussian distribution. The PDF in the previous subsection is that of the Gaussian distribution. The two parameters that define the Gaussian distribution are μ and σ. Confusion often crops up as a result of outdated statistical terminology. The **Gaussian** distribution is often referred to as the **normal** distribution. However, that does not mean that the Gaussian distribution is what everything should follow- it is just one possibility of many.

- $f(x) = \dfrac{1}{\sigma\sqrt{2\pi}} \exp(-\dfrac{(x-\mu)^2}{2\sigma^2})$ for $-\infty \le x \le \infty$

and its CDF is

- $F(X) = \dfrac{1}{\sigma\sqrt{2\pi}} \displaystyle\int_{-\infty}^{X} \exp(-\dfrac{(x-\mu)^2}{2\sigma^2})dx$

Let's return to the GSL monthy lake level record. The values for μ and σ are estimated from the histogram plotted in Fig 3.7 and a Gaussian (normal) distribution is then calculated using the sample mean and variance. Visually, you should be able to tell that the Gaussian fit in this instance is not particularly good, since the lake level is skewed (i.e., there are a few events of

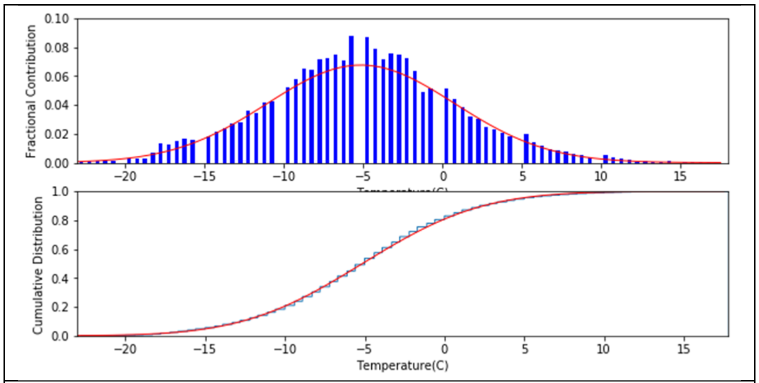

Figure 3.8. Gaussian fit to temperature at Alta in terms of its histogram (top) and cumulative distribution (bottom).

high water levels that would not be expected given the typical values of lake level and its spread about the sample mean). Also, there are fewer low water years than expected from the Gaussian distribution.

Let's examine the hourly temperature values at Collins (CLN) near Alta during winter from 1998-2005 as shown in Fig. 3.8. Although the Gaussian distribution underestimates the occurrence of temperature near the mean value, it appears that Collins winter temperature can be approximated by a Gaussian parametric distribution defined by the sample mean and variance. Note the occasional gaps in the histogram- the original data is in 1°F intervals, so there are some 0.5°C bins with no values.

Now, let's return to generic Gaussian distributions. Every variable can be transformed into standardized anomalies with mean 0 and variance 1. The leftmost panel of Fig. 3.9 indicates that for an environmental variable for which the Gaussian is a good fit to its empirical PDF, then 68.3% of the total variance is within 1 standard deviation of the mean. The middle figure indicates that 95.5% of the total variance is within 2 standard deviations of the mean while the right figure defines that 2.3% of the time we would expect that a variable explained by a Gaussian distribution would be larger than 2 standard deviations of the mean. Alternatively, we can use the quantile function to determine the x values that correspond to a particular probability. For example, if we are interested in the limits corresponding to 90% of the total variance, then that is equivalent to ±1.65σ of the mean.

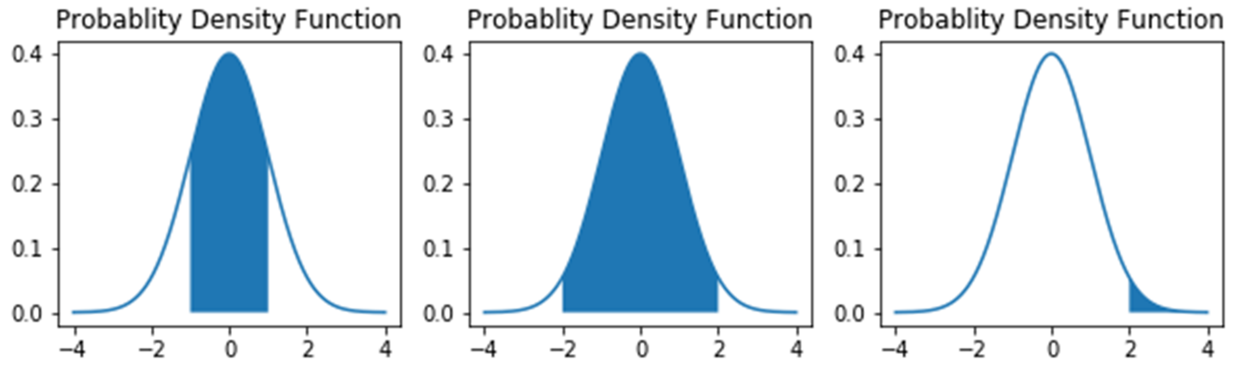

Figure 3.9. PDF's for the case when the sample mean is 0 and variance is 1. The percent values within the shaded areas are 68.3%, 95.5%, and 2.3%, respectively.

c. Other Parametric Distributions

Many environmental variables (e.g., wind speed and rainfall) are decidedly skewed to the right in part because values are nonnegative. The gamma distribution with 3 parameters is quite versatile for such situations. Other variables (e.g., wind direction, relative humidity) are constrained at both ends for which the beta distribution with 2 parameters is an appropriate choice.

Of interest for many applications, is to examine parametric distributions of extreme values, i.e., the rare events for continuous variables. There are a number of variants of theoretical distributions to describe extreme events: Gumbell, Fischer-Tippet, and Weibull, among others. However, these theoretical distributions assume random events that may not be appropriate for environmental events that often occur serially, e.g., an extreme heat wave typically will last several days in succession. If sufficient data are available, then the empirical PDF can be used to estimate the probability of rare events.

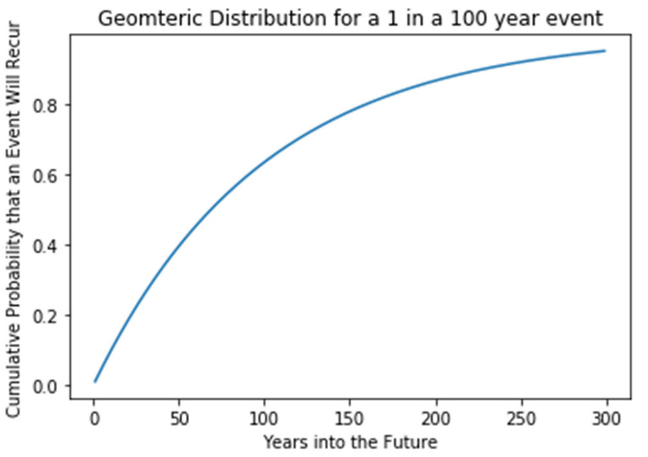

Figure 3.10. Cumulative distribution for the recurrence of a rare event- in this case a one in hundred year event assuming a geometric distribution.

Extreme values are often defined to estimate the annual probabilities of damaging events such as heavy rains or high winds. The recurrence of extreme events is frequently defined in terms of the return period, i.e., 100 year floods, etc. However, there is no guarantee that a 100-year event will happen in the next 100 years. The probability of a 1 in a 100 year random event is Pr{0.01}. The geometric distribution specifies probabilities for the number of trials required until the next success. Fig. 3.10 shows the cumulative probability of the period until the next 100 year event. In other words, if the probability of a 100-year event is 0.01, then there is only a 63% chance that it will happen in the next 100 years after the last event and there is still a 12% chance that it will not happen in 200 years. If the probability of a rare event increases to 2%, then there is a 12% chance that it will not happen in 100 years. Of course, this is only true if the event can be described by a geometric distribution.

As an example of evaluating the return period of extreme events, let’s examine the peak streamflow record from the Klamath River in northern California for the 1911 to 2020 water years (Oct.-Sept.; hence December floods are part of the following calendar year) as shown in Fig 3.11. To what extent can we estimate the occurrence of extremely high peak flows on the Klamath River by a parametric fit to the data? We have an advantage here since we can estimate empirically what a one in a hundred year event is, as we have a record of 110 years. That estimate is the red line in Fig 3.11 determined as the 99th percentile from the empirical CDF that is shown as the blue curve in the right panel of Fig. 3.12. People often estimate one in a hundred year events from records of 20 (or less) years based on parametric fits. We’ll use this example to show how that can be done, but why this approach might not give a realistic answer. Empirically in our case, the 99th percentile in the record is 523,400 ft^3/s. In this instance, there have been two “hundred year” events during the 1965 and 1974 water years.

As shown in Fig. 3.12, a Gaussian parametric fit is a poor choice in this instance to describe the peak streamflow as it would estimate many more low peak flows than observed and fewer high peak flows. The Weibull fit does a better job at capturing the skewed nature of the peak streamflow and estimates the magnitude of one in a hundred year peak streamflow fairly well (red dash-dot line in Fig. 3. 11).

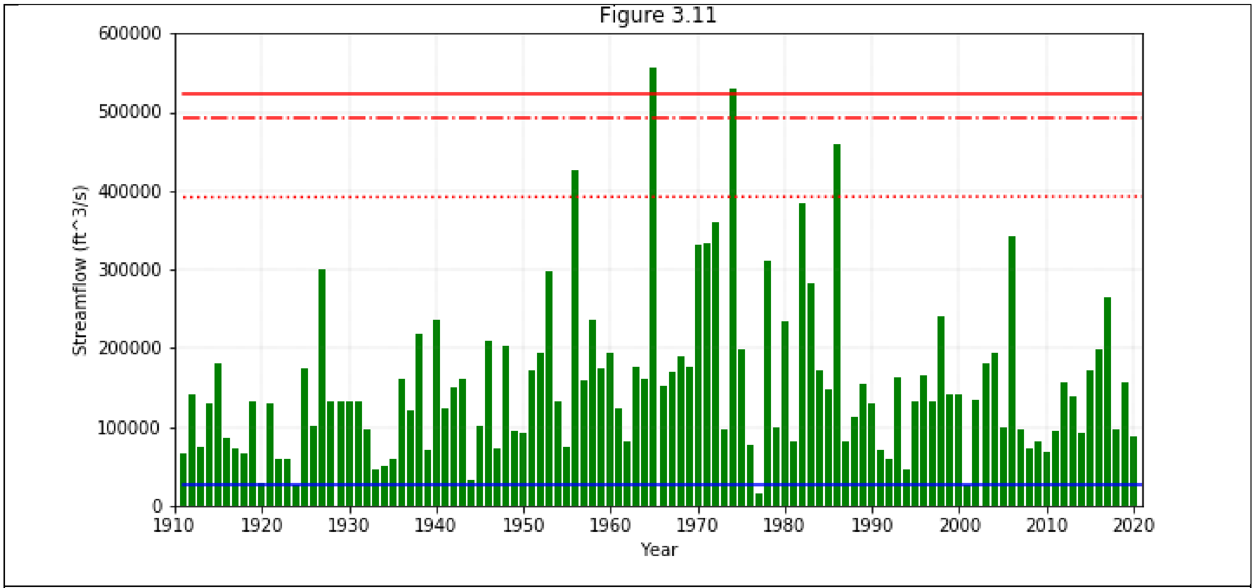

Figure 3.11. Peak streamflow (ft^3/s) during the water year for the Klamath River, CA. "100 year" peak flows (99th percentile or 1 in a 100) were observed in December 1964 (1965 water year) and 1974 (above the red line). "100 year" lowest flows were observed in 1920, 1977, and 2001 (below the blue line). The doted and dashed/dot lines indicate the 100 year peak flows estimated from Gaussian and Weibull fits respectively.

Another way of examining the "goodness" of a parametric fit is to look at probability-probability plots (Fig. 3.13). If the Weibull parametric fit was perfect, then all the blue circles (observed values) would lie along the red line. So, the Weibull fit is quite good.

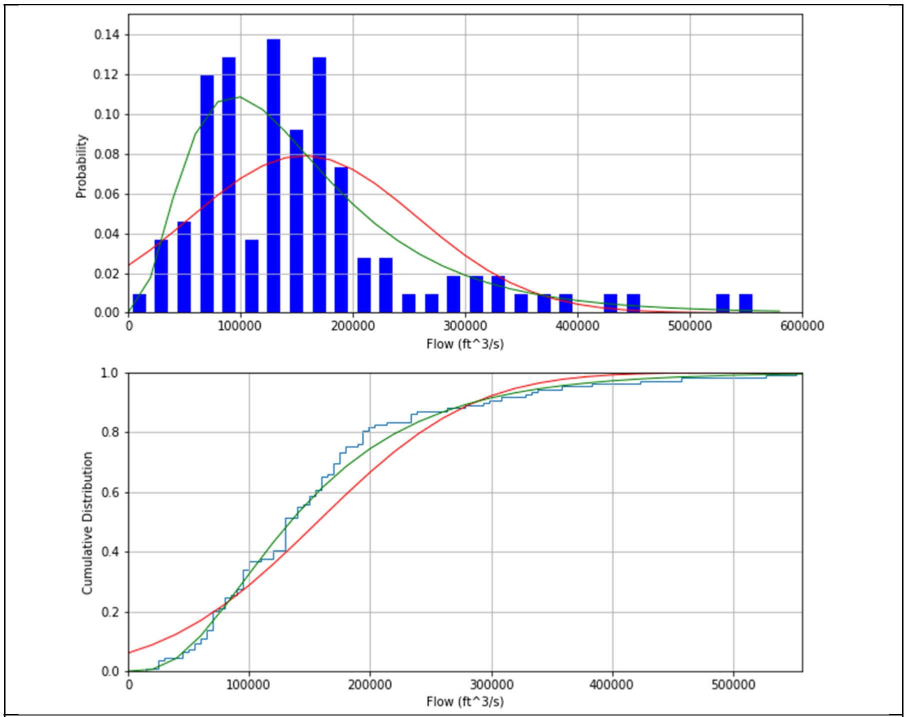

Figure 3.12. Empirical histogram (top) and CDF (bottom) of Klamath peak streamflow in blue. Red (green) curves denote Gaussian (Weibull) parametric fits to the data.

We could now ask something for which we don't have a long enough record- what would be peak streamflow for a one in a thousand year flood? We can extrapolate using our Weibull fit and guestimate that it might be 708,000 ft^3/s. That would certainly be devastating in that river corridor if it were to happen.

d. Hypothesis Testing of Means

Let's return to the annual precipitation in Utah and

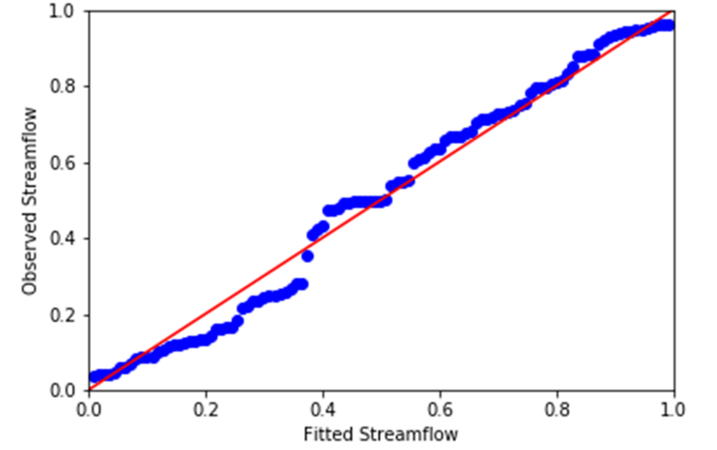

Figure 3.13. Probability-probability plot with observed Klamath peak streamflow in blue. The red straight line shows where the observed values should lie if the Weibull parametric fit the data perfectly.

use a completely arbitrary definition of a drought: that the average annual precipitation anomaly over a 5 year period differs substantively from zero. We will evaluate strings of the 5 year periods and try to objectively compare each five year period to the others.

One expectation might be that the mean precipitation anomaly during any of the 5 year periods is 0 - this would be the **null hypothesis.** The null hypothesis, H_0, defines a frame of reference against which to judge an alternative hypothesis, H_A, which in this instance could be "the mean precipitation anomaly during the past five years is less than zero".

The steps required for a hypothesis test are:

- identify a test statistic that is appropriate to the data and question at hand. The test statistic is computed from the sample data values. In this example, the 5-year sample mean will be the test statistic, but we'll also need to use the sample variance as well.
- Define a null hypothesis that we hope to reject. In this case, the null hypothesis is that the sample mean is 0.
- Define an alternative hypothesis. In this case, the sample mean is negative.
- Estimate the null distribution, which is the sampling distribution of the test statistic if the null hypothesis is true. It is very important to recognize that we need to know the sampling properties of the test statistic. That is, the sample mean could be drawn from a Gaussian parametric distribution, another parametric distribution or even we could define the sampling distribution of the mean empirically by randomly sampling over and over taking five years within the past 126 years.
- Compare the observed test statistic (the composite mean value of each 5-year period to the null distribution. Either:
    - the null hypothesis is rejected as too unlikely to have been true if the test statistic falls in an improbable region of the null distribution, i.e., the probability that the test statistic has that particular value in the null distribution is small, or,
    - the null hypothesis is not rejected since the test statistic falls within the values that are relatively common to the null distribution.

Not rejecting H_0 does not mean that the null hypothesis is true; rather, there is insufficient evidence to reject H_0. The null hypothesis is rejected if the probability, p, of the observed test statistic in the null distribution is less than or equal to a specified significance (or rejection) level denoted as the α level. Usually, 1% or 5% significance levels are used, i.e., if the odds of the test statistic occurring in the null distribution are less than 1% or 5%, then we often reject the null hypothesis. Depending on how the alternative hypothesis is framed, rejecting the null hypothesis may be equivalent to accepting the alternative hypothesis; however, there may be many possible alternative hypotheses. The first step of any significance testing is to set an appropriate α level to reject the null hypothesis. In other words, you must first set a threshold, such as 1% that denotes a 1 in 100 chance that you are accepting the risk of rejecting the null hypothesis incorrectly. This 1% risk is a Type I category error of a false rejection of the null hypothesis.

Confused? Yes, it can be difficult to make sense of this and we'll discuss that this approach is in many respects fundamentally flawed. However, statistical methods require some sense of how realistic they are and this is a flawed, but standard approach.

e. *Central Limit Theorem and Student-t Test*

Now we consider one of the reasons the Gaussian distribution is used so much. First, roll 1 six-sided die 10,000 times. That's a population as shown in the top panel of Fig. 3.14. The histogram of the population indicates that the chance of getting any one number from 1-6 is basically the same in that population (but they are not identical odds).
show Now roll 6 dice at the same time 10,000 times (bottom panels of Fig. 3.16). In other words, we have a population of 10,000 samples of 6 events, so there are now 60,000 values. We can determine each 6-member sample's sum, mean, or variance separately. The most common sum is around 21 The most common mean is around 3.5. The odds of getting a total count of 6 or 36 are small. Note that we end up with a Gaussian distribution. *The central limit theorem states that the sum (or mean) of a sample (6 dice) will have a Gaussian distribution even if the original distribution (one die) does*

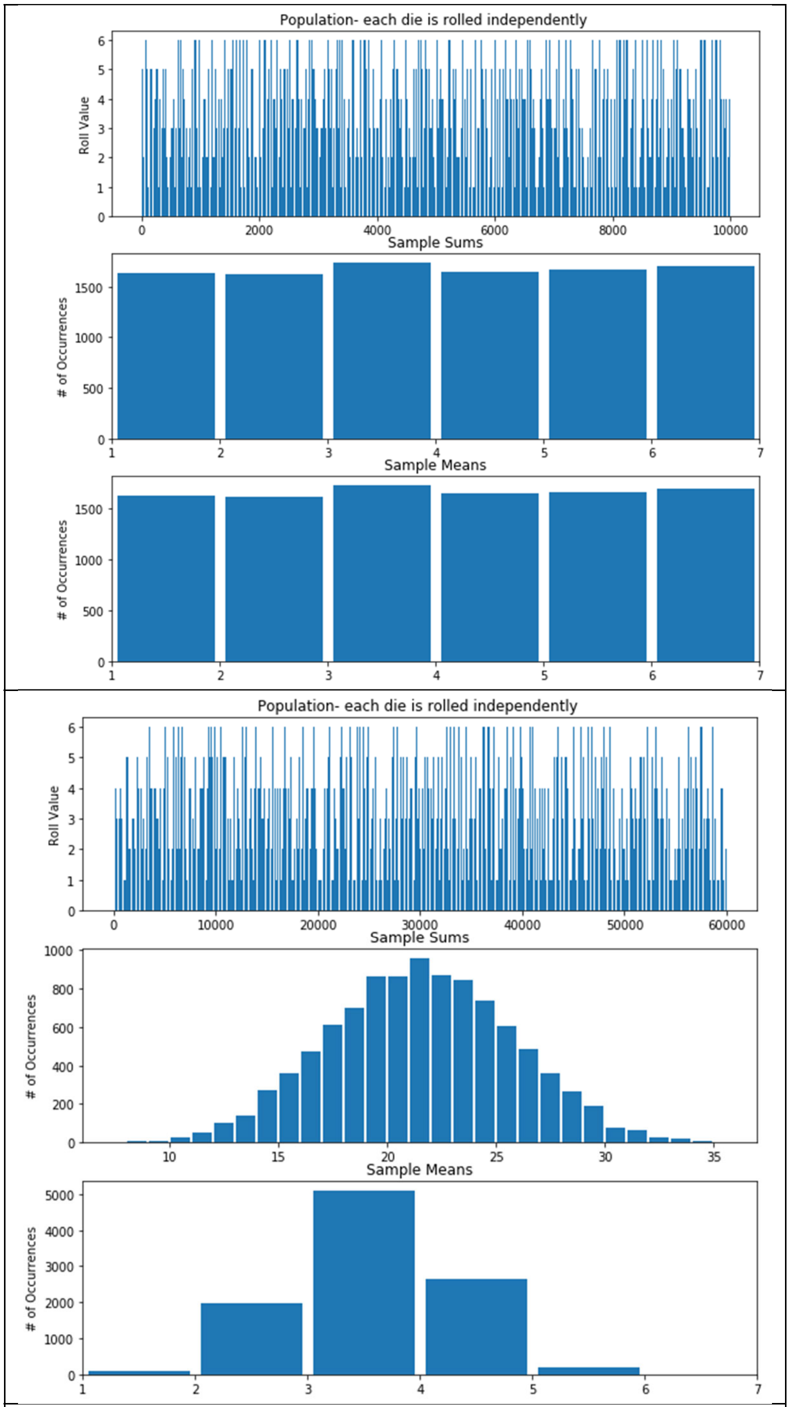

Figure 3.14. Top panels. Population and histograms s based on rolling 1 die 10000 times. Bottom panels. Population and histograms based on rolling 6 dice 10000 times.

not have a Gaussian distribution, especially as the sample size increases. In other words, $\sigma_{\bar{x}} = \sigma / \sqrt{n}$ where $\sigma_{\bar{x}}$ is the standard deviation of the sample means, σ is the standard deviation of the original population, and n is the sample size. For a large population, such as in this example, the standard deviation of the population is roughly 1.7 and that of the 6 member samples drawn from the population is roughly 0.7, which is what should be expected.

Let's return to our example attempting to determine 5-year drought periods. The sample standard deviation is 6.5 cm for the annual precipitation over the 126 years. We could randomly obtain the Gaussian distribution shown in the left panel of Fig. 3.15 with a standard deviation of 6.5 cm about the anomaly mean of 0. There is a 95% chance that the precipitation anomaly will lie between ±12.7 cm. There is a 2.5% chance that a random year could have a precipitation anomaly below -12.7 cm. So, by chance, if we had a 100 year sample, we would expect 2-3 of those years to possibly have precipitation anomalies less than that threshold.

We now randomly take 5 values and average them. If we selected 5 years at random from the population many times, then according to the central limit theorem, we'd end up with the right panel. There is a 95% chance that the 5-year sample mean would lie between ±5.7 cm of a 2.5% chance that a random sample of five years would have a precipitation anomaly less than -5.7 cm. In other words, it becomes easier at least in terms of the magnitude of a single value vs mean of 5 values to have an extreme 5-year mean ("a drought" according to this lame definition) than just to have one extreme dry year.

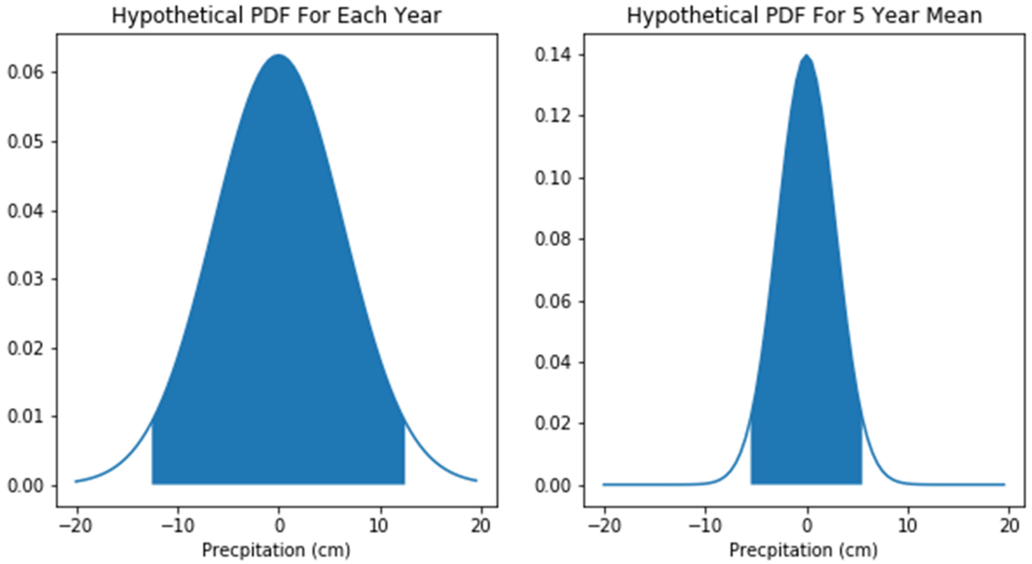

Figure 3.15. Gaussian distribution with standard deviation equal to 6.4 cm (left panel) and 2.9 cm (right panel).

We use the central limit theorem as a way to determine whether a mean from a particular sample differs **significantly** from the mean we specify as being appropriate for the null hypothesis assuming that we know something about the population variance. Assume for the moment that the *population* standard deviation was 6.5 cm as assumed in the left panel of Fig. 3.15. In the last 5 years (2016-2020), the annual precipitation anomalies are 1.5, -1.5, -4.6, 9.0.-15.6. The mean anomaly over those 5 years is only -2.2 cm, even with the super low precipitation in 2020. Then we would determine that we could NOT reject the null hypothesis at the 2.5% level (the left-tail of the distribution) , since the sample mean during the last 5 years of -2.2 cm first lies within the shaded area in the right panel of Fig. 3.15 (the right unshaded portion would be shaded in too).

If we go back to 1900-1904 (centered on 1902) when the precipitation departures were -10.0, -5.7, -8.2, -5.3, -2.9 cm, then the 5-year average is -6.4 cm, which is lower than the -5.7 cm limit associated with our 2.5% threshold to reject the null hypothesis. We can state that in this instance we can reject the null hypothesis that a negative 5-year mean differs from 0 at the 2.5% level. However, we just “cherry-picked” a case- did I have any reason ahead of time to look at the 1900-1904 period? NO- I ran the analysis and then went searching for one of the cases that meet my definition of “significance”- that is an aposteriori approach (after the fact). We could expect

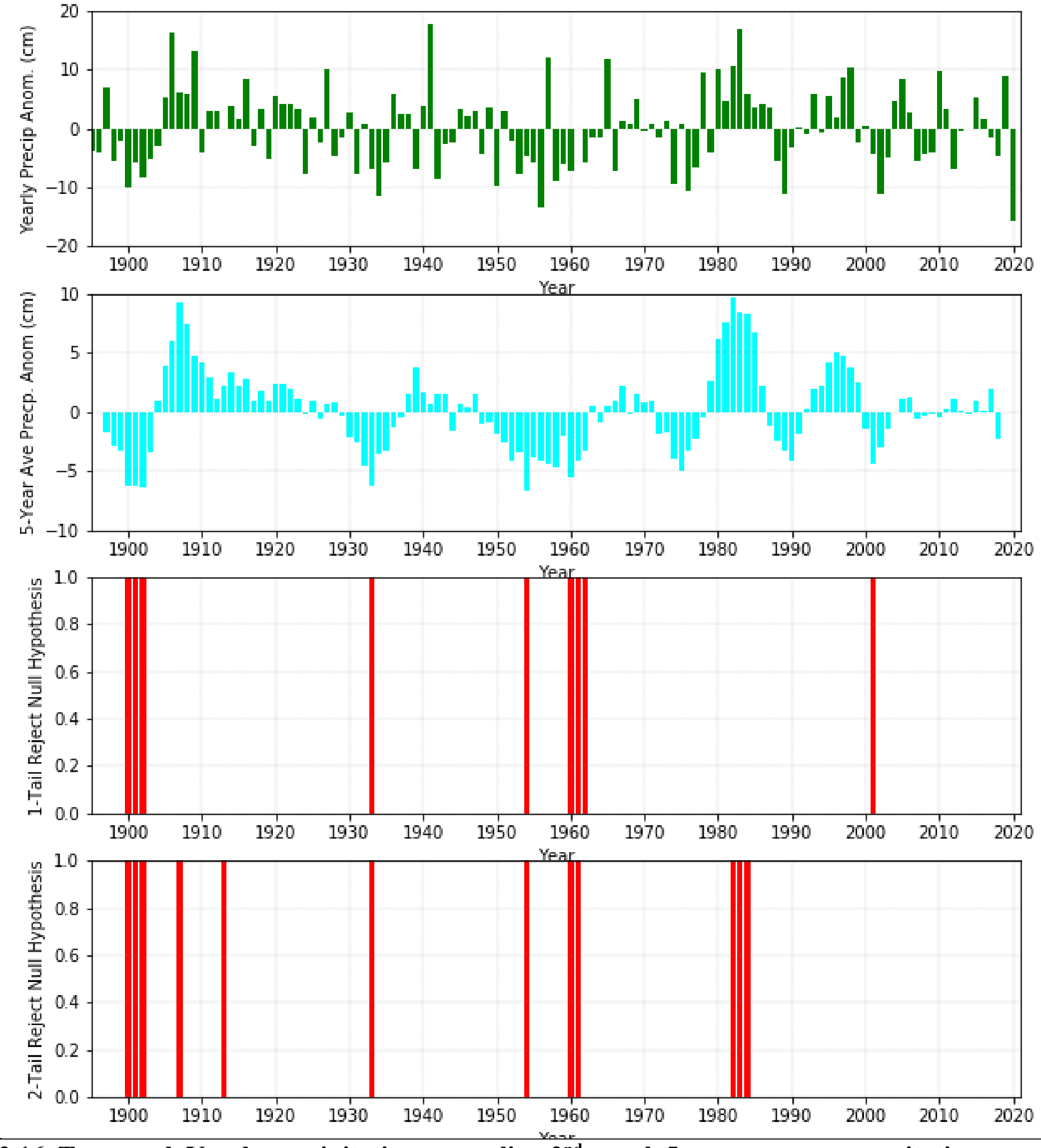

Figure 3.16. Top panel. Yearly precipitation anomalies. 2nd panel. 5-year average precpitation anomalies. 3rd panel. Values of 1 indicate the null hypothesis can be rejected at the 5% confidence level for a 1-sided left t-test. Bottom panel. Values of 1 indicate the null hypothesis can be rejected at the 5% confidence level for a 2-sided t-test (both wet and dry events).

in a 122 year sample of 5-year means (we can't compute the 5-year means for the two years on the ends of the time series) that there would be ~3 such years to possibly happen by chance.

The above is not the correct way to assess the situation for small sample sizes, as we didn't factor in the sample size explicitly. Usually we only have an *estimate* of the population variance from our sample. Then, as already discussed in Chapter 2, the sample standard deviation $s_x = \sqrt{\frac{n-1}{n}}\sigma$ or $\sigma_{\bar{x}} = s_x / \sqrt{n-1}$. The degrees of freedom is n-1, which is a reminder that the sample can be described by the mean (1 value) plus n-1 others.

The **Student's t test** is a way to determine whether the null hypothesis can be rejected. The name "Student's t" comes from an employee of the Guiness brewery who had to submit his paper as "Student" anonymously to a journal. The t value is defined as: $t = (\bar{x} - \mu)\sqrt{n-1} / s_x$, which can be shown to be normally distributed for large numbers of degrees of freedom (n-1 greater than 30 or so). So, the students-t test is a conventional way to adjust when using a smaller sample size, such as our five years for this drought definition.

There are a variety of ways to grasp the meaning of the t statistic. The simplest is to visualize the numerator as the 'signal', the difference between the sample and null hypothesis means times the number of members of the sample, and the denominator as the 'noise', the variability within the sample. As the value of t gets larger, our confidence in rejecting the null hypothesis that the mean of the sample is zero gets higher. The t value is large if: (1) the spread between the sample mean and the null value mean is large, (2) the number of members in the sample is large, (3) the variability in the sample is small.

Let's loop over all 5-year samples in our record to see which periods might be considered droughts. The top panel of Fig. 3.16 shows the yearly precipitation anomalies. Note that the precipitation anomalies of -13.5 and -15.6 cm in 1955 and 2020, respectively are the only ones that would be considered exceptional and for which we might be able to reject the null hypothesis for a single year with 2.5% confidence if this was a random sample drawn from a Gaussian distribution as shown in the left panel of Fig. 3.15.

We want to know which 5-year periods can be classified as droughts and have some confidence that calling them a drought is not just due to chance. The 2nd panel shows the 5-year averages, so the differences from zero is the numerator for the t-test. The denominator for the t-test can be guestimated from the top panel (or look at the values for each year in the code) as to the variability of the 5 yearly values centered on the middle year.

We first set our **rejection limit, α**, the probability level for rejecting the null hypothesis. If we are only interested in droughts, then if we set α = 5%, we are asking in our case: is the random probability that the t-statistic for a particular sample of 5 years lower than 5% and is the t-value negative? This is a one-tail t-test (the left tail). If so, then we can reject the null hypothesis accepting a 5% risk that it might happen by chance. The center panel highlights the middle year of the 5 years when the null hypothesis can be rejected, i.e., accepting a 5% risk in those situations that classifying them as droughts could simply have happened by chance. The 5-year periods where it is not possible to reject the null hypothesis are blank in the third panel. If we are willing to accept a higher risk of falsely rejecting the null hypothesis, then we could use a higher threshold of say 10% and thereby identify more drought episodes. Note that our 1900-1902 period is identified as being one of the periods for which we can reject the null hypothesis but even with the abnormally low precipitation in 2020, the 5-year mean from 2016-2020 doesn't meet our objective criteria for a drought.

Now for another caveat- we had 122 opportunities to reject the null hypothesis. So, 5% of 122 is ~6, which means we might expect purely by chance that we would have 6 drought periods lasting 5 years. We found 9 (but there were really only 5 independent events), so we are teetering on the edge of not really finding much useful about drought periods in Utah. As a general rule, researchers tend to ignore these types of issues related to serial dependence in time series when assessing statistical significance.

So far in this simple example, the test of the sample mean is a one-sided (or tailed) 'left' test (we're only interested in droughts). A two-sided test would require an alternative hypothesis that the 5-year mean anomaly is simply nonzero (either positive or negative) akin to what was discussed in relation to Fig. 3.15. We're now interested in both "droughts" and "wet" periods- 5-year periods when the average is greater than zero. This alternative hypothesis implies that any of the 5-year mean values must be even further from 0 (a smaller p value), i.e., a 2.5% chance for drought periods and a 2.5% chance for wet periods. It may seem somewhat paradoxical that assuming an alternative hypothesis (both wet and dry) leads to a tougher obstacle to reject the null hypothesis but that is due to keeping the rejection limit, α, at 5%. This becomes evident in the bottom panel of Fig 3.16 since the null hypothesis can be rejected for only a smaller number of really strong "drought" periods. But, now we can reject the null hypothesis for a number of 5-year wet periods, including the early 1980's. Remember that the "signal" (the mean of the 5 values) is evaluated relative to the "noise", the variance within the sample. So, we are more likely to be able to reject the null hypothesis when the variability within the sample is small. Note that we now have 7 "independent" periods of either low or high precipitation in the state, only one more than what we could expect from chance. *(The python code can be confusing. The scipy module provides a two-tailed test. The one-sided test is easily computed by simply dividing the probability returned from the two-tailed test by two and considering the sign of the t value. Dividing the probability by two makes it easier to reject the null hypothesis, the reasons for which are described here.)*

f. Summary

The exploratory data techniques developed in Chapter 2 are simply that: exploratory. Research involves defining a testable hypothesis and demonstrating that any statistical test of that hypothesis meets basic standards. Typical failings of many studies include: (1) ignoring serial correlation in environmental time series that reduces the estimates of the number of degrees of freedom and (2) ignoring spatial correlation in environmental fields that increases the number of trials that are being determined simultaneously. The latter inflates the opportunities for the null hypothesis to be rejected falsely. Use common sense. Be very conservative in estimating the degrees of freedom temporally and spatially. Avoid attributing confidence to a desired result when similar relationships are showing up far removed from your area of interest for no obvious reason. The best methods for testing a hypothesis rely heavily on independent evaluation using additional data not used in the original statistical analysis. We'll introduce those concepts in the next chapter.

4 Exploratory Multivariate Data Analysis

The techniques in Chapter 2 focus on a sample of data from a single variable independent of others. An advantage of many environmental data sets is the large number of simultaneous measurements available. We often want to relate how one or more phenomena are related to others. Besides simply measuring different quantities, we often have access to observations at different locations (both horizontally and vertically). Hence, our sample may have many dimensions: x, y, z, t, and variable, model, etc. The number of dimensions can easily grow beyond that. For example, if we are dealing with forecasts, then the forecast lead time or perturbations of model parameterizations or initial conditions become other dimensions. Dealing with the dimensionality of environmental data sets in statistical analyses is of general concern (see Murphy 1991; *Mon. Wea. Rev.,* 1590-1601). Obviously, we can slice such data sets up in a number of different ways to simplify the dimensionality of the problem depending on the goals of the study. Exploratory multivariate data analysis encompasses an array of tools to assess relationships between two or more samples.

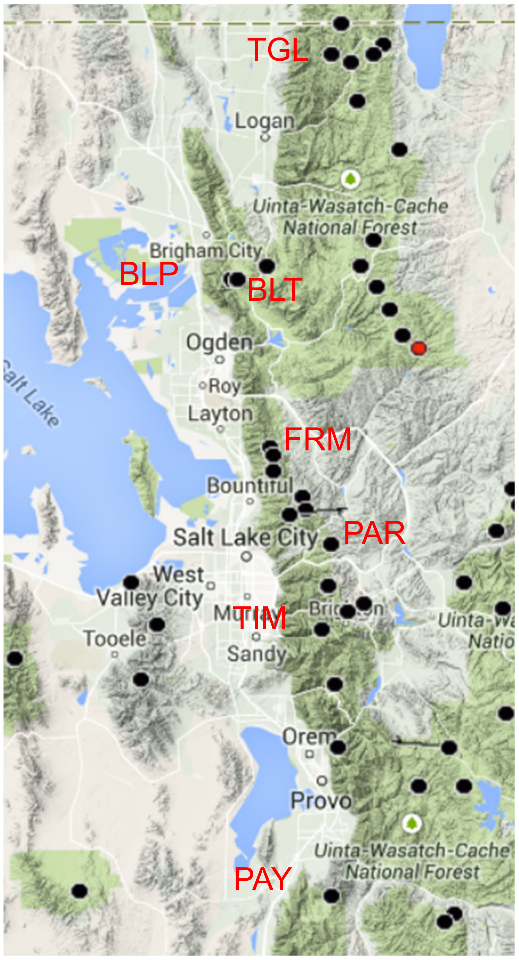

Figure 4.1. Locations of the 7 SNOTEL sites examined

a. *Linear Regression Between Two Variates*

We'll use a data set of monthly precipitation collected at high elevation (SNOTEL) sites in the Wasatch Mountains. To keep the analysis manageable, only the time series of precipitation at the 7 stations labeled in Fig. 4.1 will be used and the data are preprocessed in the code to consider only the water year (October-September) totals. These data can be obtained from this link.

The top panel of Figure 4.2 shows the time series of total precipitation at Ben Lomond Peak and Ben Lomond Trail over a 41-year period. Since the stations are very close to one another, it is not surprising that the year-to-year variations in precipitation at the two sites are very similar. However, since Ben Lomond Trail is at a lower elevation, then its precipitation is distinctly less than that at Ben Lomond Peak. The degree of similarity within the two pairs of time series is easier to evaluate after transforming the data into standardized anomalies (bottom panel of Fig. 4.2). You might expect that if we try to estimate the precipitation at Ben Lomond Peak from that at Ben Lomond Trail we should be able to do well. You should also recognize that the degrees of freedom in these records is fewer than the 41 years in the sample, maybe something like 20?

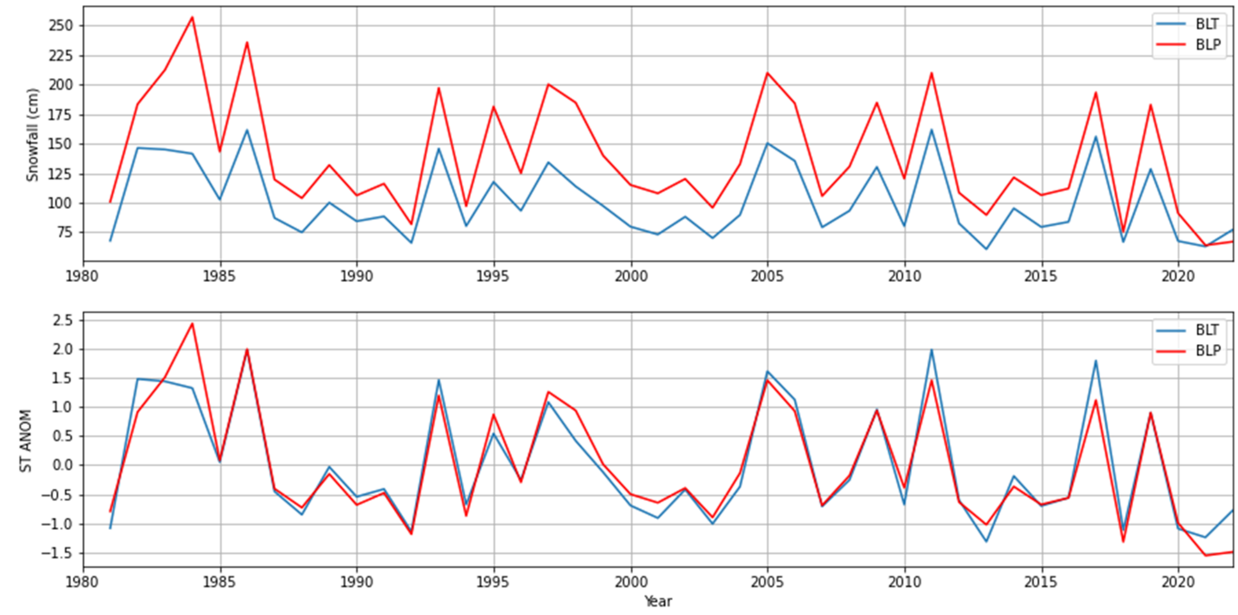

Figure 4.2. Time series (cm) of seasonal total precipitation at Ben Lomond Peak and Trail (top panel) and standardized anomalies (nondimensional) of the time series in the lower panel.

Scatter plots of the values associated with two variables are a convenient way to examine relationships between paired data. Clustering, spread, outliers, etc. become apparent in scatter plots. Scatter plots can be done in terms of the raw values, anomalies, or standardized anomalies depending on the application. Since temporal continuity is lost when looking at scatter plots generated from time series of data, you need to be careful to not simply assume that each pair of observations is independent of the others.

Figure 4.3 shows scatter plots of the original and standardized anomalies for the Ben Lomond time series. The meaning of the lines in each of the panels will become apparent below. Scatter plots are easier to interpret when there is a clear one-to-one association between the two variables, i.e., for a given value of x, the values of y in the sample are similar to one another. If the scatter plot looks like a blob, then that is a clear indication of a lack of one-to-one association. If the pairs of values tend to fall along a line, then it is appropriate to think of the two variables as being linearly related to one another. They may instead exhibit quadratic or higher order association. The scatter plots between the Ben Lomond stations reflect linear correspondence but the greater precipitation at the Peak in 1984 compared to the Trail suggests less association for that year.

As part of exploratory data analysis it is common to want to estimate the values of one variable from that of another. I'm going to avoid saying 'predict' one variable from the other for the moment. Let's start by trying to estimate precipitation at Ben Lomond Peak from the values at Ben Lomond Trail. First, we know that there is more precipitation on average at the higher elevation site, so we need to consider the differences between the two means. The simplest linear approach is to assume that for a given value at Ben Lomond Trail $\hat{x}_i$, our estimate $\hat{y}_i$ (where the subscript *i* refers to a particular year) at Ben Lomond Peak can be determined as follows:

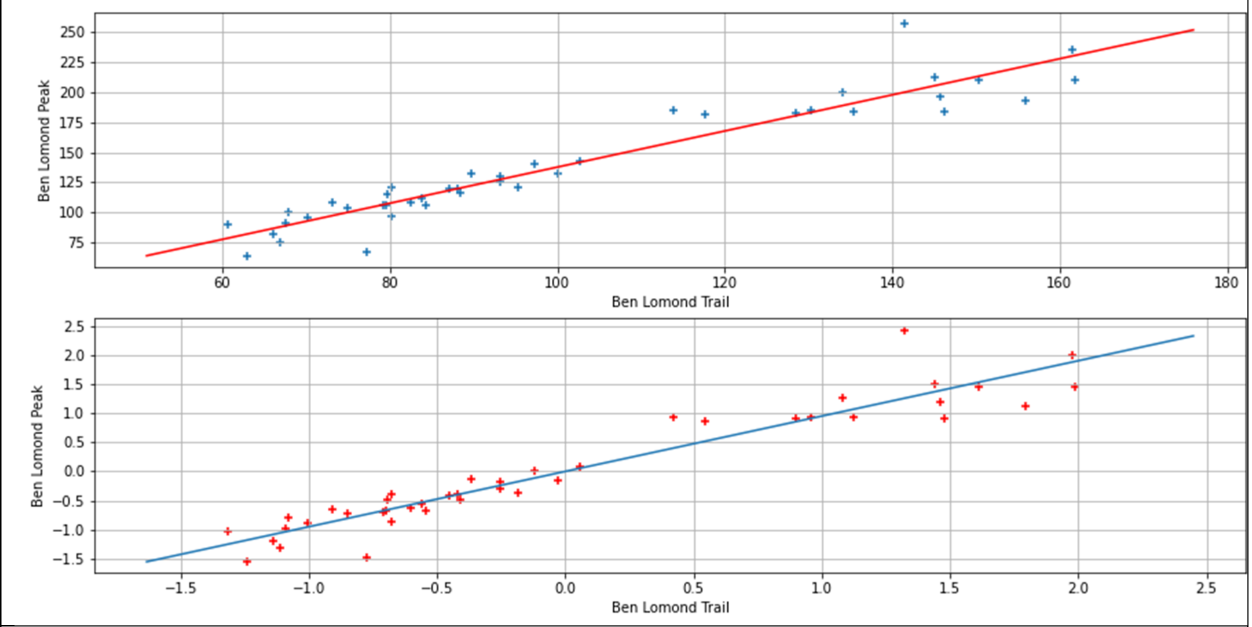

Figure 4.3. Scatter plot of total precipitation (cm) at Ben Lomond Peak vs. Trail (top panel) and of their standardized anomalies (bottom panel).

$\hat{y}_i = \bar{y} + b(\hat{x}_i - \bar{x})$ *(4a.1)*

That relationship takes into consideration the differences in the two means. We obviously need to figure out how to determine the coefficient, b, and one approach is to use our sample of data collected over the 40 year period. First, consider the red line in the top panel of Fig. 4.3. It is a particular linear estimate using a specific value of b. If the Ben Lomond Trail precipitation is 140 cm, then we would estimate Ben Lomond Peak to measure ~200 cm.

Alternatively, we can use another coefficient, r, to estimate the standardized anomalies at Ben Lomond Peak from the standardized anomalies at Ben Lomond Trail as$\hat{y}^*{}_i = r\hat{x}^*{}_i$ where the asterisk indicates a standardized anomaly and *r* again needs to be determined from the pairs of values in the samples. A linear estimate for a particular value of *r* is shown by the line in the lower panel of Fig. 4.3 If *r* is 1, then the standardized anomalies at the two sites would be estimated to be exactly the same. If *r* is -1, then they would have the same magnitudes but opposite signs of anomalies. If *r* is 0, then for any x standardized anomaly, the estimate for y would be 0.

How good are those estimates? We can use our sample of data to compute the errors for these specific choices of b (the slope of the line). For example, we have several observations of Ben Lomond Trail precipitation between 140 and 150 cm and during those years, Ben Lomond Peak measured between 180 and 250 cm. Obviously, our linear estimate didn't do particularly well in two of those cases, but most of the other years had closer estimates to those observed.

Any particular error in the estimate can be written as $e_i = y_i^{'} - \hat{y}_i$, which is the distance between the line and the specific observation. The best line will be the one which minimizes all the distances e_i, so we want $\sum_{i=1}^{n} e_i^2$to be a minimum. For our sample values $y_i' = bx_i' + e_i$, where the primes denote deviations from the respective means. Then if we use the entire sample:

$\overline{y_i'^2} = b^2\overline{x_i'^2} + 2b\overline{x_i'e_i} + \overline{e_i^2}$ *(4.a.2)*

The term on the left is the sample variance of y about the mean and is given as the sum of the variance explained by the linear fit + how the errors and the deviations of x are related over the entire sample + the variance that is not explained by the linear fit, which is what we want to be small. The middle term on the right is assumed to be zero, because e_i is assumed to be random if our sample is large enough.

- Then $s_y{}^2 = b^2 s_x{}^2 + \overline{e_i^2}$ *(4.a.3)*

We want to choose b so that the explained variance of the linear fit (the first term on the right) is as big as possible and the last term is as small as possible.

To minimize $\sum_{i=1}^{n} e_i^2$ means to determine $\frac{\partial}{\partial b}\sum_{i=1}^{n} e_i^2 = 0$, which by substituting in for e_i yields

$$\frac{\partial}{\partial b}\sum_{i=1}^{n} e_i^2 = \frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i' - bx_i')^2 = 2\sum_{i=1}^{n}(y_i' - bx_i')(-x_i') = 0$$

or $\sum_{i=1}^{n} x_i'y_i' = b\sum_{i=1}^{n}(x_i')^2$. Dividing through by n, using the definition for a mean, and rearranging yields

- $b = \overline{x_i'y_i'}/\overline{(x_i')^2} = \overline{x_i'y_i'}/s_x^2$ *(4.a.4)*

where $\overline{x_i'y_i'}$ is called the covariance and relates how departures from the mean of x and y are related. The covariance has units of the quantity squared, like a variance. Covariances are used in many disciplines: turbulence, planetary-scale dynamics, etc. The covariance is:

- large and positive if there is a general tendency in the sample for large and positive (and/or negative) anomalies of x occurring when large positive (negative) anomalies of y are observed
- large and negative when there is a general tendency for large and positive (and/or negative) anomalies of x to occur at the same time as large negative (positive) anomalies of y when aggregated over the entire sample
- near zero when there is a general tendency for cancellation within the sample, i.e., sometimes large positive anomaly values of x are associated with large positive anomaly values of y and other times large positive anomaly values of x are associated with large negative anomaly values of y.

Returning to *4.a.2*, and dividing through by y's sample variance, then we have: $1 = \frac{b^2 s_x{}^2}{s_y{}^2} + \frac{\overline{e_i^2}}{s_y{}^2}$, which simply says that a fraction of y's variance is due to the variance estimated by our linear regression estimate and the remaining fraction is due to random (or unexplained) errors. Defining r^2 as the squared linear correlation coefficient:

- $r^2 = b^2 s_x{}^2/s_y{}^2 = (\overline{x_i'y_i'})^2/(s_x{}^2 s_y{}^2)$ *(4.a.5)* or
- $r = (\overline{x_i'y_i'})/\sqrt{\overline{x_i'^2\, y_i'^2}}$ -1≤r≤1 *(4.a.6)*

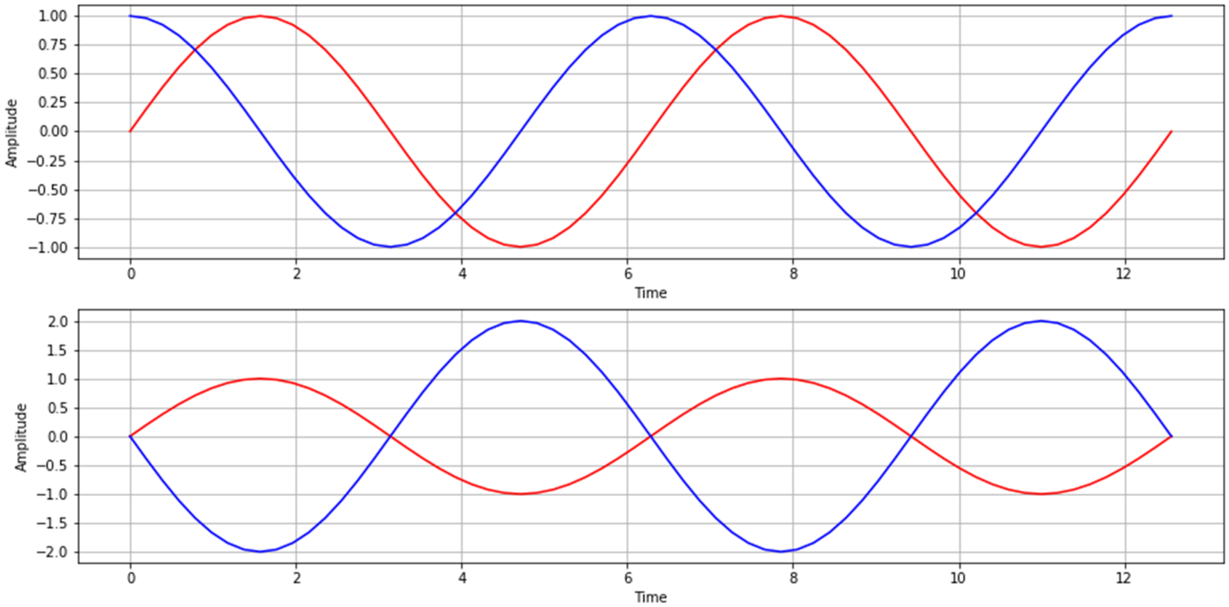

Figure 4.4. The linear correlation between the top two time series is 0 while that between the lower two time series is -1.

In addition, if we standardize the anomalies of x and y by dividing the anomalies by their respective standard deviations:

- $x_i^* = x_i' / s_x , y_i^* = y_i' / s_y , r = (\overline{x_i^* y_i^*})$ *(4.a.7)*

Then, y's sample variance can be described alternatively as: $1 = r^2 + \frac{\overline{e_i^2}}{{s_y}^2}$where the squared correlation coefficient is the fraction of the total variance of y estimated from x. While the covariance is not bounded, $-1 \leq r \leq 1$and when:

- r = 1 - the linear fit estimates all of the variability of the y anomalies and the standardized anomalies of x and y vary identically
- r = -1 - the linear fit estimates all of the variability of the y anomalies in the sample but when the standardized x anomaly is positive, then the standardized y anomaly is negative
- r = 0 - the linear fit explains none of the variability of the y anomalies in the sample and the standardized anomalies of x and y have no relationship to one another in the sample.

If r=0, then the only thing we can say is that the best linear estimation for y is its mean value. If the scatter plot looks like a blob, then the linear correlation coefficient is likely to be close to zero, as there is no linear fit to the data that is going to explain any of the variability of y. As r approaches 1 (or -1), then we gain confidence that we can estimate the behavior of the second variable from the first, and vice versa. The squared correlation coefficient defines the fraction of variance that the two variables have "in common".

The coefficients b and r can be computed using several different approaches. One approach is that the sums of the product $x_i y_i$ be computed as well as the sum of squares and sums of the two variables. i.e., $cov = \overline{xy} - \overline{xy}$. This is a useful approach when processing large data sets. The second approach uses linear algebra. Define the column vector $\vec{X}'$ for the x anomalies (BLT) and the column vector $\vec{Y}'$ for the y anomalies (BLP or TGL), then

$\vec{X}' = \begin{bmatrix} x_1' \\ x_2' \\ \dots \\ x_n' \end{bmatrix}$ and $\vec{Y}' = \begin{bmatrix} y_1' \\ y_2' \\ \dots \\ y_n' \end{bmatrix}$ and the covariance $\overline{x_i'y_i'} = \vec{X}'^T\vec{Y}'/n$ *(3.a.7)*

where the superscript T denotes the transpose of the column vector (i.e., the column vector is switched to a row vector). The resulting matrix multiplication of the 1 x n row vector times the n x 1 column vector yields a scalar number, which divided by the total number of elements, is the average of the vector product. Similar matrix multiplications can be done to obtain the sample variances.

The linear fits are as shown in the above figures and the linear correlation between the precipitation anomalies at Ben Lomond Peak and Trail is 0.96, which is really high. Hence, 91% of the variance of total precipitation at Ben Lomond Peak can be explained by the variability of total precipitation at Ben Lomond Trail and only 9% of the variance is unexplained.

The **Pearson** correlation coefficient is another name for the linear correlation coefficient defined here. The Pearson correlation coefficient is not a robust and reliant statistical measure, because the covariance and variance terms are quite sensitive to outliers. The **Spearman** rank correlation coefficient is a more robust measure and it is determined by sorting the data for the two variables in order from least to greatest and then computing the covariance as a function of rank, i.e., the correlation would be high if the highest (and lowest) values occur at the same time in both records. The Spearman approach is particularly appropriate for analyzing variables with skewed distributions, e.g., precipitation and wind speed. In our simple example of the correlation between Ben Lomond Peak and Trail, the Pearson and Spearman correlations are identical.

There are a number of limitations of linear correlation coefficients that must be recognized:

- There is a widespread tendency to use correlation coefficients of 0.5-0.6 to be indicators of "useful" association. However, 75%-64% of the total variance is unexplained by a linear relationship if the correlation is in that range.
- Linear correlations can be made large by leaving in signals that may be irrelevant to the analysis. For example, if we correlate over many years two temperature records from opposite sides of the earth, the linear correlation will be large if we do not remove the annual cycle. Perhaps we may be interested in knowing that the annual cycle in Great Britain is similar to that in North Dakota, but usually we are more interested in examining departures from the seasonal cycle.
- Large linear correlations between two variates may occur simply at random, especially if we try to correlate one variate with many, many others. This situation arises frequently when we relate interannual or intraseasonal anomalies in one part of the globe to those over the entire globe. Tests are available to weed out some of these situations. We will formalize later what steps should be taken when an unexpected strong association crops up vs. one that we have hypothesized to exist.
- Relationships in the data that are inherently nonlinear will not be handled well.
- When two time series are in quadrature with one another (e.g., one time series corresponds to a cosine and another corresponds to a sine) as shown in Fig. 4.4, then the linear correlation is 0. You should be able to recognize that as the relative phase of two

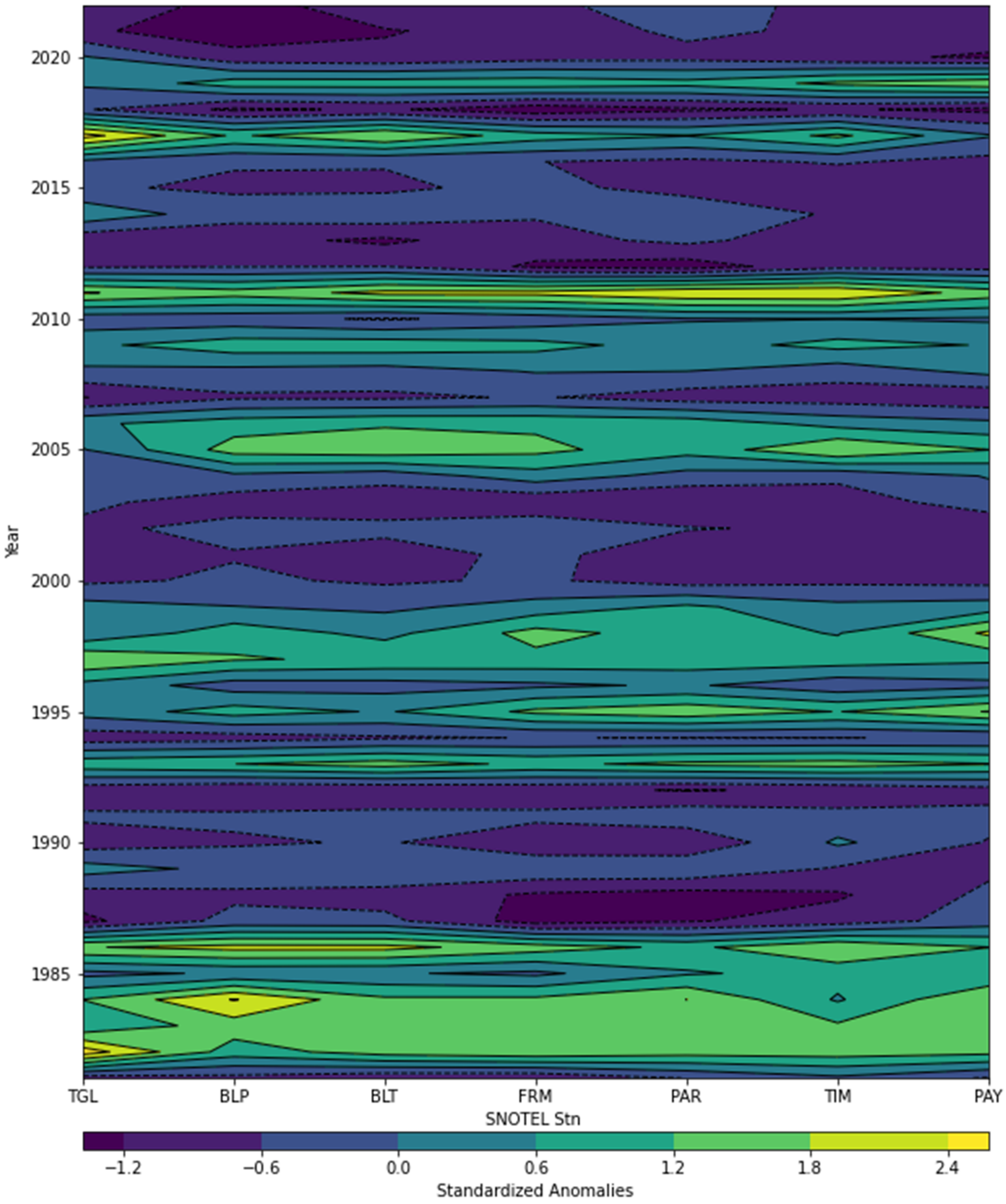

Figure 4.5. Hovmuller diagram (time decreasing down the page and location advancing south across the page) of standardized precipitation anomalies.

sinusoidal time series progresses from 0 to 90 to 180, then the linear correlation changes from 1 to 0 to -1. Since the environment is filled with propagating features, the limitations of the use of linear correlations for such phenomena should be readily apparent.

- Linear correlation provides no information on the relative amplitudes of two time series. For example, the linear correlation between the lower two time series in Fig. 4.4. is -1.0, yet the amplitude of one of the time series is 2 times that of the other. The normalization by the standard deviation of each variable removes the relative amplitude information.

b. Multivariate Linear Correlations

As an extension to exploratory tools for pairs of data, it is straightforward to simultaneously examine the association between many simultaneous observations such as the water year totals at the 7 sites highlighted in Fig. 4.1.

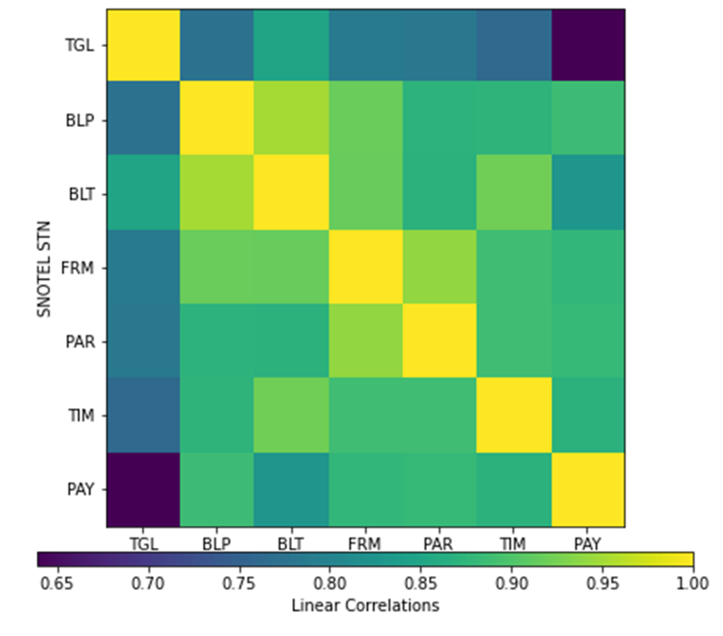

Figure 4.6. Correlation between the 7 pairs of SNOTEL precipitation time series computed over the 42 year sample.

We can compute the average and sample standard deviation for each of the 7 stations over all 42 years (n=42) and thereby computed the standardized anomalies for each station as a function of time. Then, we can define the n x 7 two-dimensional array of standardized anomalies as $\vec{X}*$ where n is the total number of years and 7 is the number of stations. In other words,

$$\vec{X}*=\begin{bmatrix} x*_{11} & x*_{12} & \dots & x*_{17} \\ x*_{21} & x*_{22} & \dots & x*_{27} \\ \dots & \dots & \dots & \dots \\ x*_{n1} & x*_{n2} & \dots & x*_{n7} \end{bmatrix} \quad (4.b.1)$$

A Hovmuller diagram (time vs. location) is simply a plot of the matrix defined in 4.b.1. For example in Fig. 4.5, nearly all the stations show similar year-to-year variations, but there are some differences. For example, all the stations had large standardized precipitation anomalies during the 2005 season but the positive standardized anomaly at Tony Grove (TGL) was smaller than that at all the other locations during that year. Tony Grove had its largest precipitation anomaly during 1982. All stations reported low precipitation amounts in 2021.

If all the observations are loaded into a single dataframe then the linear correlations between all pairs of simultaneous observations can be calculated as shown in Fig. 4.6. The diagonal elements are 1.0, as they are the correlations of the 7 time series with themselves. The correlation matrix is symmetric, i.e., the values are the same for each row/column pair.

There are obviously some strong linear associations among all 7. The weakest is between the two stations furthest apart: TGL and PAY and equal to 0.64. In other words, those two time series explain 40% of the variance of the other, and 60% is not explained by their co-linear variations. This result shouldn’t be too surprising, since they are the ones separated by the largest distance and there are some differences in the temporal evolution of the precipitation anomalies over time evident in the Hovmuller diagram of Fig. 4.5. TGL is less correlated overall with the 6 others. We could use the Spearman correlation to reduce the sensitivity of the correlation matrix to outliers. The differences are trivial in this case.

The above exploration of the data centers on the question: how do the yearly precipitation departures from the 42-year temporal mean at one location compare to those at another location when considered over all 42 years?

Linear correlations of this sort are commonplace in environmental fields. A time series of one variable is often correlated with time series of variables at every location on a grid. That results in a temporal anomaly correlation map. "**Teleconnection**" maps are where the time series at each gridpoint of a variable are related to the time series of that same variable at every point and then this procedure is repeated for every possible gridpoint. An example of that is shown in the top panel of Fig. 4.7 generated by creating a time series from Create time series: https://psl.noaa.gov/data/correlation/custom.html for 500 mb height for January from 1948-2022 at a gridpoint near Salt Lake City (40N, 112.5W). Here is the final link used to generate the time series and here is the link to generate the plot.

Nearby time series of 500 mb height anomalies near Salt Lake City are positively correlated with the time series above Salt Lake City (and the correlation is 1.0 at the gridpoint we started from). What is of interest are the relatively strong negative correlations upstream (<-0.5 to the west) and downstream (<-0.4 to the east) of Salt Lake City that put together exhibit a wavelike pattern. When there is anomalous ridging aloft (few storms in January) affecting Salt Lake City, then there is a deep trough in the Gulf of Alaska (lots of storms). Alternatively, lots of ridging in the Gulf of Alaska tends to be associated with more troughs aloft (storms) in Salt Lake City and the western U.S. What about the weak linear correlations further away from Salt Lake City?

Anomaly correlation maps with many different climate indices can be computed from the CDC web site: Correlations: https://psl.noaa.gov/data/correlation/. For example, the middle panel of Fig. 4.7 shows the correlation between an index referred to as the Pacific North American (PNA) index that is derived as a simple weighted average of 500 mb height in specific locations (near Hawaii, in the

Gulf of Alaska, western Canada, and the southeastern US). The PNA index was derived decades ago as an indicator of a leading monthly/seasonal weather pattern in the Northern Hemisphere https://www.cpc.ncep.noaa.gov/data/teledoc/pna.shtml . Note that 500 mb height anomalies near Salt Lake City are not strongly correlated to the PNA index in January. Here is the link to generate the figure.

The lower panel of Fig. 4.7 shows the linear correlation between 500 mb height anomalies in the Northern Hemisphere and sea surface temperature anomalies in the equatorial Pacific (for regions referred to as regions 3 and 4). Positive correlations imply that when the SST in the equatorial Pacific is above (below) normal then 500 mb heights are above (below) normal. The tendency during El Nino winters for enhanced troughing (lower than normal 500 mb heights) in the Gulf of Alaska and over the southern U.S. combined with above normal heights in western Canada is evident. However, the linear correlation between equatorial SST and 500 mb height during January in the vicinity of Salt Lake City is close to zero. Here is the link to generate that figure.

Note that all three correlation patterns in Fig. 4.7 are similar, but not identical. Which ones are "significant" and which aren't? We'll tackle that later.

The above analysis has focused on how the year-to-year variations in precipitation (rows) at locations (columns) relate to similar variations at other locations. Alternatively, we could transpose the original matrix and view the data as elements of maps (m rows) at specific times (n columns):

$$\hat{\hat{X}} = \begin{bmatrix} \hat{x}_{1,1} & \hat{x}_{1,2} & \dots & \hat{x}_{1,n} \\ \hat{x}_{2,1} & \hat{x}_{2,2} & \dots & \hat{x}_{2,n} \\ \dots & \dots & \dots & \dots \\ \hat{x}_{m,1} & \hat{x}_{m,2} & \dots & \hat{x}_{m,n} \end{bmatrix} \quad (4.b.3)$$

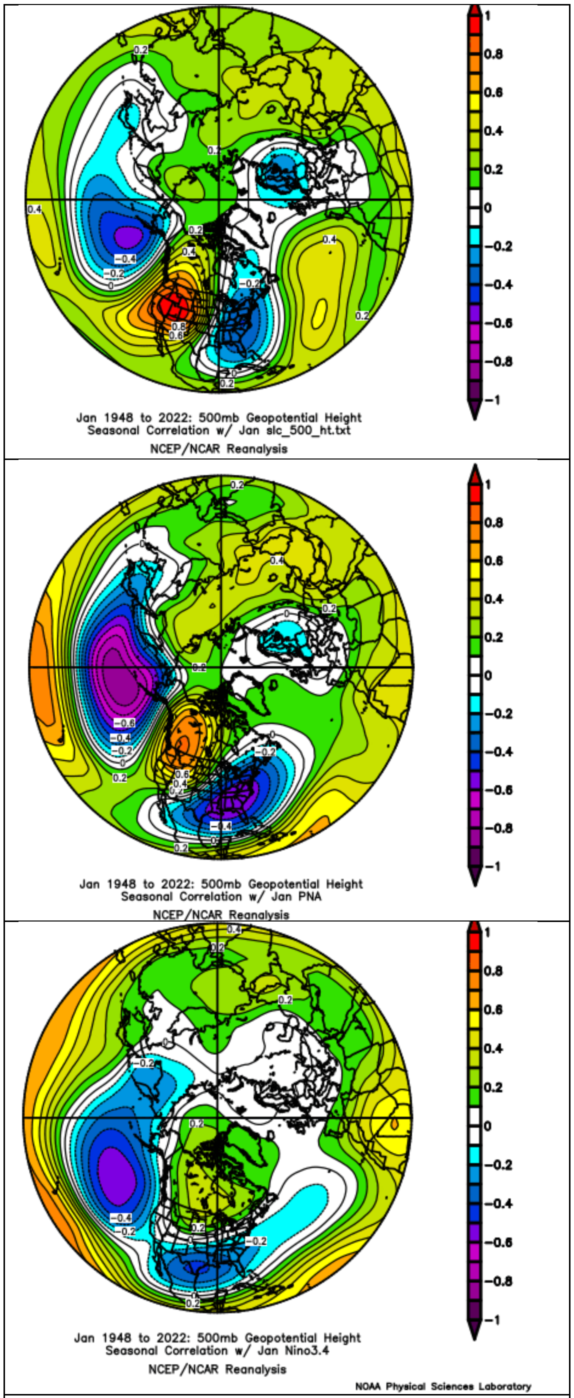
Jan 1948 to 2022: 500mb Geopotential Height
Seasonal Correlation w/ Jan slc_500_ht.txt
NCEP/NCAR Reanalysis
Jan 1948 to 2022: 500mb Geopotential Height
Seasonal Correlation w/ Jan PNA
NCEP/NCAR Reanalysis
Jan 1948 to 2022: 500mb Geopotential Height
Seasonal Correlation w/ Jan Nino3.4
NCEP/NCAR Reanalysis
NOAA Physical Sciences Laboratory
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Figure 4.7. Top. Correlation between time series of January 500 mb height throughout the NH and 500 mb height near Salt Lake City. Middle. Same but with the PNA index. Bottom. Same but with the equatorial Pacific SST index.

We can then compute the spatial average over the locations or map elements and the variability about that spatial average for a time. We can compute the linear correlation coefficients between every pair of maps from

- $\vec{S} = \hat{\vec{X}}*^T\,\hat{\vec{X}}* \quad /m$, *(4.b.4)*

Linear correlations between pairs of anomaly maps are commonly used to verify model forecast fields vs. analysis grids. Usually, the long-term daily mean is removed at each grid point and then the departures from the spatial mean are computed for the forecast and analysis grids. Such spatial anomaly correlations have been computed for forecast grids by the operational centers as shown in Fig. 4.8 (in this case for the 5-day and 10-day 500 mb height forecast grids in the Northern and Hemisphere from NCEP). If the spatial anomaly correlation was equal to one, then the forecast and the analysis would exhibit the same spatial anomaly patterns. If the correlation is 0, then the model forecast and analysis fields are completely unrelated in a linear sense.

These spatial anomaly correlations between two fields are computed as follows. Let the analysis grids at m locations (rows) and n times (columns) be $\hat{\vec{X}}'$ and the forecast grids for one specific model at m locations and n times be $\hat{\vec{Y}}'$. Then we can compute the spatial anomaly correlations between every matched pair of forecast and verifying analysis maps and generate a figure like that from

- $\vec{S} = \hat{\vec{X}}'^T\hat{\vec{Y}}'/m$, *(4.b.5)*

Besides the information on the relative accuracy of the various models shown in Fig. 4.8, the magnitude of the anomaly correlations indicates greater accuracy in the Northern Hemisphere at 5 days compared to 10-day lead time. All of the caveats regarding linear correlation apply to the spatial anomaly correlations. Hence, for this type of forecast verification, we are unable to assess if the forecasts have large errors in amplitude. In addition, a relatively good forecast with a slight phasing error (i.e., ridges and troughs captured properly but shifted in longitude) will be counted as a relatively poor forecast. For many other examples of the uses of spatial anomaly correlations and other accuracy measures, browse

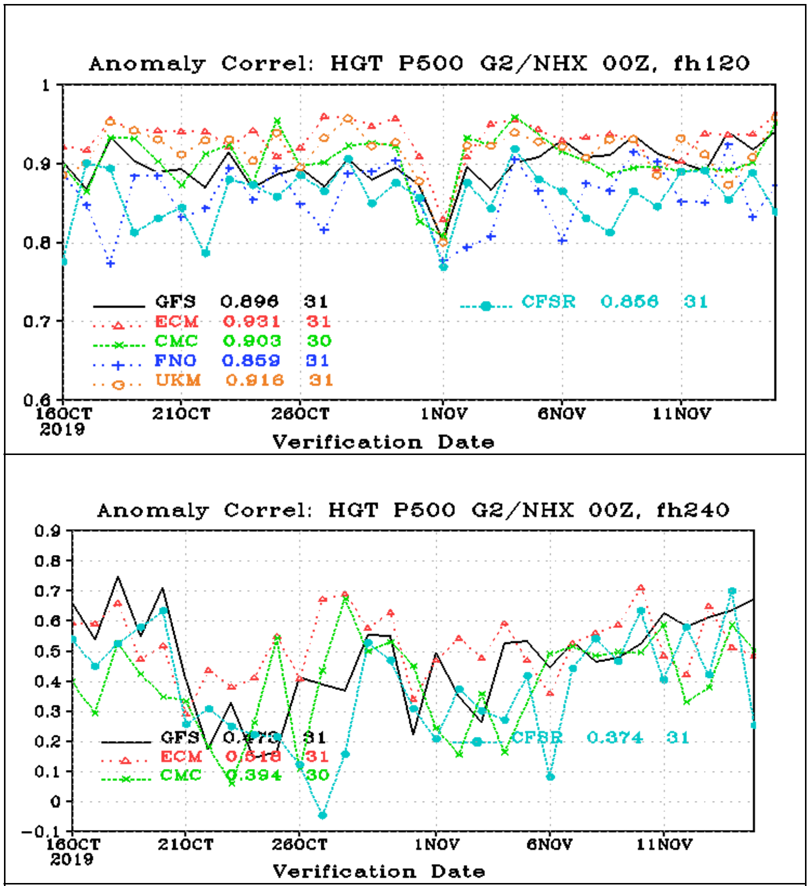

Figure 4.8. Anomaly correlations between 120 h (top) and 240 h (bottom) model forecasts and analyses computed each day.

around https://www.emc. ncep.noaa.gov/ gmb/STATS_vsdb

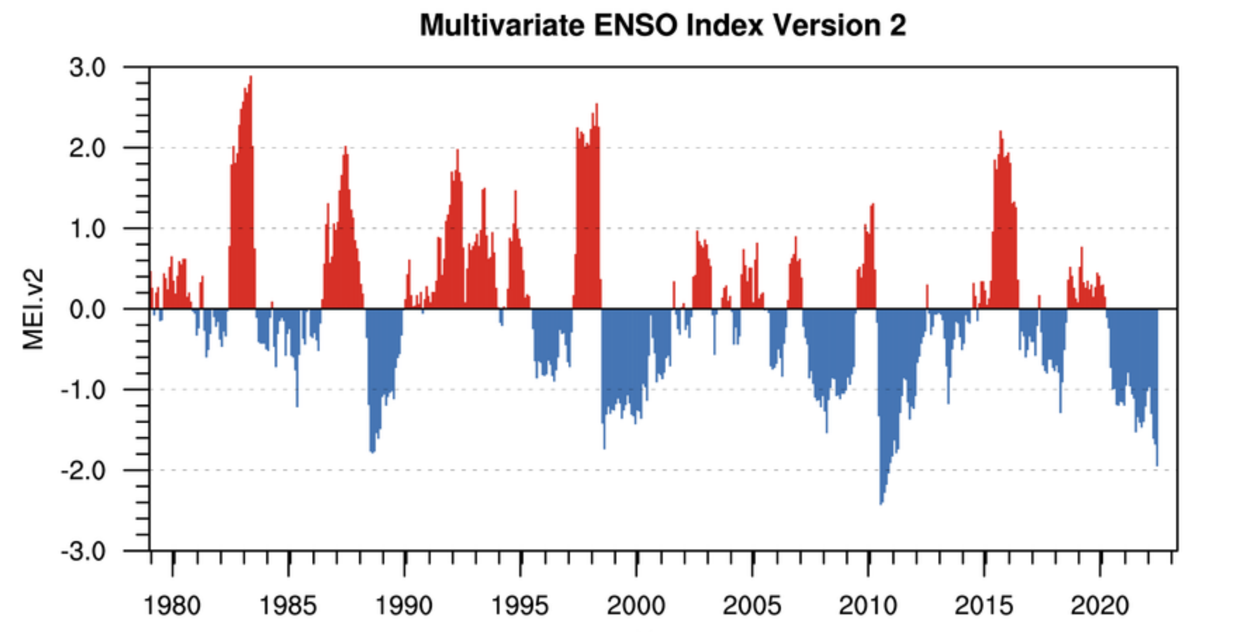

Figure 4.9. Multivariate ENSO index

c. Compositing

Compositing (or superposed epoch analysis) is frequently used to assess the common environmental features associated with a sample of events. For example, the occurrences of some relatively rare event are identified (e.g., local floods or warm sea surface temperature in the equatorial Pacific). The goal is to identify the average conditions within some large data set before, during, and after those rare events. The availability of the NCEP/NCAR reanalysis grids and the CDC web software available at https://psl.noaa.gov/data/composites/day/ and https://psl.noaa.gov/cgi-bin/data/composites/printpage.pl has helped to spawn a cottage industry of compositing applications.

The steps in the compositing process can be summarized as follows:

- select the basis for compositing and define the categories on which the compositing will be defined. It is preferable to have some physical reasoning for the categories or else the results may have limited usefulness.
- compute the means and statistics for each category
- organize and display the results
- validate the results (the methods for which we will discuss later) either in terms of: significance tests; breaking the data record into parts and showing that the results are reproducible in smaller samples; examining the relationship on an independent data set; show consistency in space and time; or verify consistency with a well-founded theory.

Relating environmental phenomena to El Nino/Southern Oscillation (ENSO) variability is of interest in many fields. The multivariate ENSO index (Fig. 4.9) is one of the better indicators of ENSO variability(https://psl.noaa.gov/enso/mei/). It is possible to identify when the biggest El Nino and La Nina events have occurred. We have been in La Nina conditions of late. For this example, I'll limit it to just the top 6 years during Jan-Feb during the available period of record for the MEI: 1983, 1987, 1992, 2010, 2016.

Using https://psl.noaa.gov/cgi-bin/data/composites/printpage.pl it is very straightforward to develop the composite 500 mb height anomaly map shown in Fig. 4.11 for those 6 January's. While the basic information obtained from this simple composite is similar to that obtained from the linear correlation shown in Fig. 4.8 between the MEI and 500 mb height anomalies (i.e., below normal heights in the Gulf of Alaska and over the southern United States), the composite analysis provides information on the amplitude of the anomalies as well.

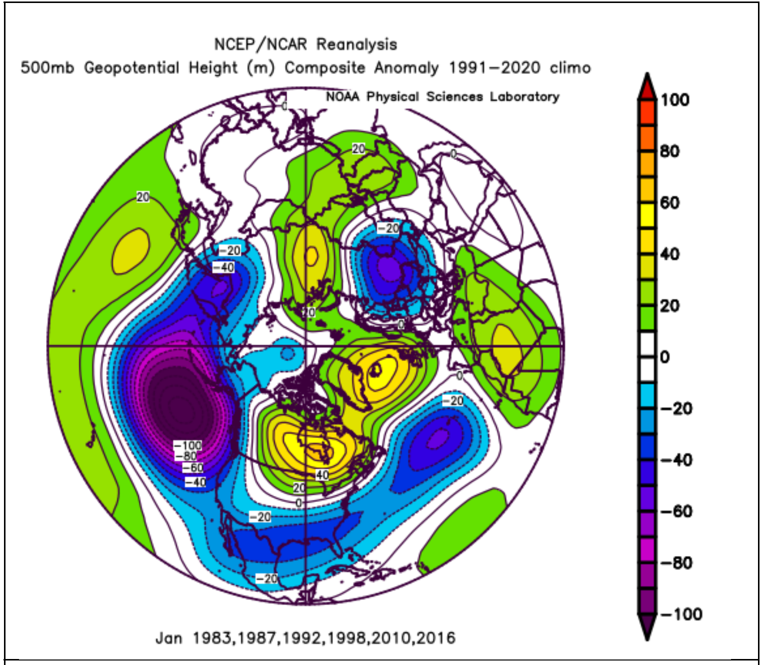
NCEP/NCAR Reanalysis
500mb Geopotential Height (m) Composite Anomaly 1991–2020 climo
NOAA Physical Sciences Laboratory
100
80
60
40
20
0
−20
−40
−60
−80
−100
-20
-40
20
-100
-80
-60
-40
40
0
Jan 1983,1987,1992,1998,2010,2016

Figure 4.11. Composite 500 mb height anomaly for the 6 January's with the highest MEI index.

One of the principal strengths of composite analysis relative to linear correlation analysis is that no assumption about the linearity of the system is made in the composite analysis. As will be discussed later, the primary limitation on composite analysis is the extent to which the sample mean can be judged to differ from the population mean. That will depend on the sample size and how much variability is present within the members of the sample.

Compositing studies need to be carefully evaluated:

- was there a reason before the analysis started to expect the relationship found in the study? We will discuss the advantage of *a priori* expectations in greater detail later.
- what is the basis for choosing the compositing categories? How arbitrary was the selection or is it based on physical reasoning?
- was there an opportunity for subjective judgment or bias to enter the composite analysis?
- do the composite results make sense logically and physically? Are there simpler explanations possible?

d. Enhancing Confidence Using Cross Validation

An approach to assess the confidence of linear regression results obtained from a sample is to apply the linear regression to an independent sample. The data could be divided in half at the outset and only half is used to "train" the regression while the rest is kept for the verification of the regression. How the data are split between the "dependent/training" and "independent/testing" samples can be tricky, especially if there are long term trends or other systematic behavior within the data set. Bootstrapping approaches are a super-repetitive approach where the data are sampled hundred of times, leaving out as few as one or several values in order to determine if there is uncertainty arising from a few outlier cases.

The Chapter 4 code includes a simple example where Payson snowfall is to be estimated from the Tony Grove snowfall. That relationship based on the 42 year sample was ok (linear correlation of 0.64 or ~40% of the variance shared in common). However, the root-mean-squared error (RMSE) between the Payson predicted values estimated by the black line in Fig. 4.12 relative to the actual snowfall (all the red and green dots) is 11.7 cm, meaning typically the annual totals estimate at Payson are off by several inches. Errors are particularly large when Tony Grove Snowfall is in the 120-140 cm range.

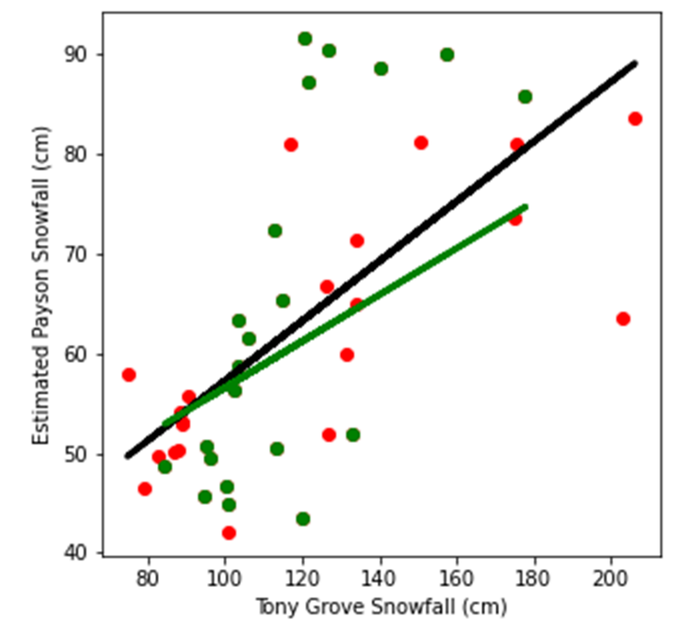

Fig 4.12. Linear fit of Payson snowfall estimated by Tony Grove snowfall using all years (the training years- red dots- and testing years-green dots) is shown by the black line. The green line is based on the linear fit using only the independent years (green dots).

How much confidence do we have in that linear prediction? We use cross validation and split the 42 years into two 21 year random samples. The red dots in Fig. 4.12 are the training (dependent) data while the green dots are the testing (independent) data. We linearly fit for the training data only, which yields the green line. Then we recompute the RMSE between the estimated values in the testing years (green dots) relative to the green line. The RMSE is now approaching 14.8 cm, larger, but not too much larger, which is also evident by the only slight different slope of the black and green lines. So, our original estimate is likely ok.

The coefficient of determination is another metric that attempts to estimate the "goodness of the linear fit". Values approaching 1 indicate a perfect estimate (what you would get if you estimated the same data twice) while approaching 0 is really bad. For the 42 year sample, the coefficient of determination was 0.41 and dropped to 0.27 for the cross-validation test. So, not a great fit, which is certainly evident in Fig. 4.12.

e. Principal Components

The year-to-year variations in precipitation at the Wasatch stations are generally similar. If we wanted one index of Wasatch mountain precipitation over time, should we just average together the 7 values each year? Could there be more than one mode of year-to-year variability tucked away in the data that we can identify? Principal component (PC) analysis is one approach to identify structures in large multivariate data sets. The purpose of PC analysis is to reduce the dimensionality of data sets through linear recombination of the variables. It is an exploratory analysis tool that can provide insight into spatial and temporal variations within a data sample. PC analysis is one of the most widely used multivariate statistical analyses in the environmental sciences, has a long history in other fields, and figures prominently in machine learning applications.

There are some key caveats about the approach:

- it depends on linear regression which suffers from the many limitations described earlier
- some of the implicit mathematical assumptions of the technique constrain the results into predictable and often artificial modes.

PC analysis is not the only approach available to you to find modes. Some of the others include:

- Canonical correlation: identify pairs of patterns in two data sets
- Principal oscillation analysis (POP): fit data to a linear low-order model
- Discriminant analysis: separate data into groups defined in advance
- Cluster analysis: separate data into groups based on similarity within sample of data
- Artificial neural networks: adaptive system that evolves as information input during learning phase
- Harmonic and spectral techniques.

Each of these alternative methods have advantages and disadvantages. Before using PCs or any other of these approaches, you need to spend some time deciding which is the most appropriate to use for your application.

There are still some critical preprocessing steps to consider:

- What are the critical temporal and spatial scales for the phenomena of interest?
- Should trends or seasonal or diurnal cycles be removed?
- Are the phenomena sampled frequently enough to observe without aliasing?
- Should the data be transformed to reduce skewness, i.e., increase normality?
- Have outliers been eliminated?

Let's assume our goal is to use PC analysis to define a single index of Wasatch precipitation based on the 7 SNOTEL time series, i.e., a reduction in the location dimension from 7 to 1. From the 2-dimensional array of standardized anomalies shown in Fig. 4.5 (the Hovmuller diagram), we can compute the linear correlation coefficients between every pair of stations (pairs of columns) as shown in Fig. 4.6. The large linear correlations between the precipitation time series evident in Fig. 4.6 imply there are fewer than 7 independent time series in the sample. So, it is not unreasonable to expect that we might be able to come up with a Wasatch precipitation index.

Define the first principal component time series to be the time series that explains the maximum ***sum*** of shared variance among all 7 time series. There are a total of 7 units of variance in the data set since each time series has been standardized to uit variance. This time series is defined in terms of a linear combination of the 7 standardized anomaly time series as follows with the year indicated by index i:

$$p_{i,1} = (e_{1,1}x'_{i,1} + e_{2,1}x'_{i,2} + e_{3,1}x'_{i,3} + e_{4,1}x'_{i,4} + e_{5,1}x'_{i,5} + e_{6,1}x'_{i,6} + e_{7,1}x'_{i,7}) / \lambda_1$$

If we start with m=7 columns, then we end up with m=7 principal components or 6 more equations like the above. Do we really need all of those time series? The goal is to have one (or a few) of these principal components explain such a large fraction of the total variance in the data set that we can ignore the rest. Now, for some matrix algebra by writing:

$$\vec{P} = \vec{X}'\vec{E}\Lambda^{-1}$$

The loadings or weights (E's) applied to the original time series are the elements of a eigenvector array (m x m) E that can be computed from the correlation array R where Λ is a diagonal eigenvalue array (m x m). Confused? You should be. The crux of PC analysis follows from the characteristics of symmetric matrices such as R. Any symmetric matrix can be decomposed into eigenvalues and eigenvectors as follows:

$$\vec{R}\vec{E} = \vec{E}\vec{\Lambda}$$

This is all handled in the Chapter 4 python code in a couple of lines. What we obtain from PC analysis are three types of output:

- Eigenvalues (the λ's) :
  - Amount of normalized variance explained by principal components
  - Percent variance explained by the first principal component is λ_1*100/m where m is the number of original time series
- Principal components (the P's):
  - If the original rows are elements of a time series, then the principal components are time series
  - These time series are linearly independent of one another (which means they are linearly uncorrelated with each other)
- Eigenvectors (the E's):
  - if the original columns were locations, then the eigenvectors are recombinations of locations, or think of them as maps
  - Each of these "maps" are linearly independent of each other (or spatially uncorrelated)

As shown in the Chapter 4 code, the first principal component explains 88% of the total variance within the 7 precipitation time series. The second explains 6% and the rest explain trivial additional amounts of the total variance. This is an example of a useful application of principal component analysis. Explaining 88% of the total variance is very good and reflects that it is reasonable to use this index to describe the temporal variability in Wasatch precipitation (and that year-to-year variations in precipitation are similar along the Wasatch). We've collapsed the seven time series into one index that is defined by a weighted combination of the 7 original time series. Those weights are determined empirically from the linear correlations between the time series and not determined from the particular values of the time series directly. As shown in the code, the first principal component index is independent of the remaining linear combinations both in time and space and the second and higher principal components explain limited variance.

Figure 4.13 shows the PC1 and PC2 time series. Compare the PC1 time series to Fig 3.5 and note that locations do have higher seasonal precipitation during 2011 and low precipitation in 2021 and 2022. Be cautious about attributing any meaning to the PC2 time series- it only explains 6% of the variance. All too often people attribute physical meaning to principal components that explain limited variance that is likely inappropriate.

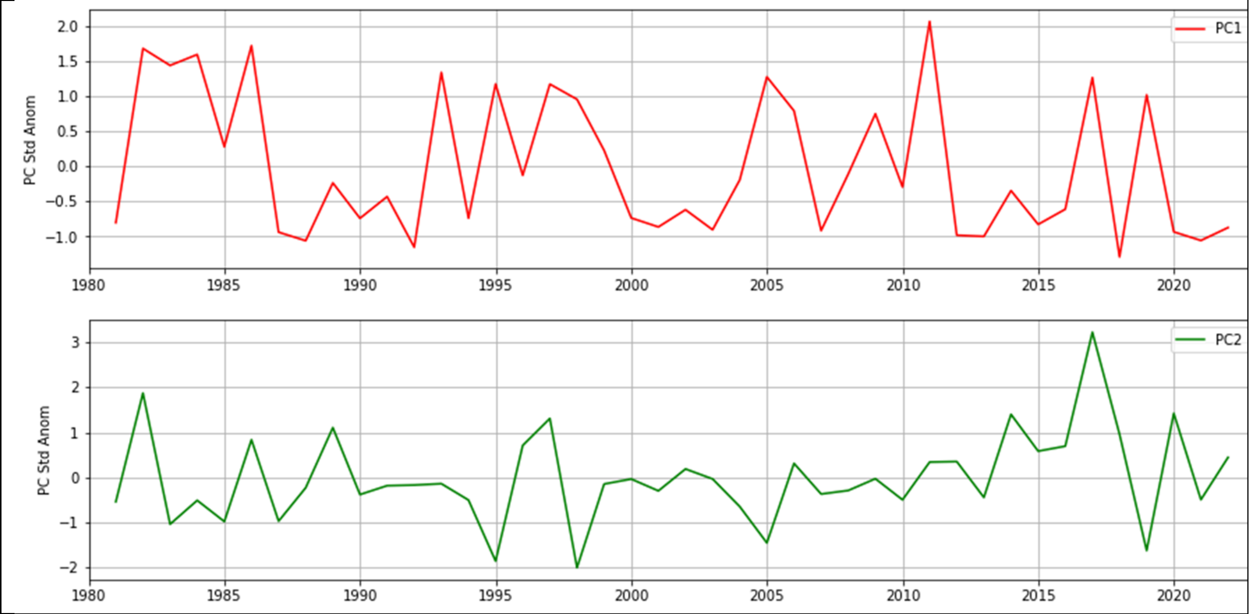

Figure 4.13. Time series of PC1 and PC2 derived from the 7 SNOTEL precipitation time series.

Figure 4.14 shows the eigenvectors, the correlations between the first two principal component time series and the 7 original time series. There are very high correlations between all 7 original time series and PC1, but the correlation between the first principal component and Tony Grove is weaker. The second principal component is poorly correlated with all of the original time series but a bit higher with Tony Grove (negatively correlated) and Payson (positively correlated).

The limitations of principal component analysis have been recognized for many years. The orthogonality (linear independence) constraint of the eigenvectors leads to artificial and predictable structures for 2nd and higher eigenvectors relative to the preceding ones. In addition, the results may change if more (or fewer) SNOTEL stations were used. Principal component analysis can be very sensitive to the domain of the analysis if the scale of any underlying modes is smaller than that of the domain.

This is only a very brief overview of PC analysis. The number of modes to consider *significant* is a sampling issue and discussed at length in the literature. One general rule is to look for a clear drop off in explained variance among the eigenvalues Another way is to only retain principal components that have 1 normalized unit of variance or more. When consecutive principal components explain comparable percentages of variance, then alternative linear combinations are equally valid for that number of modes.

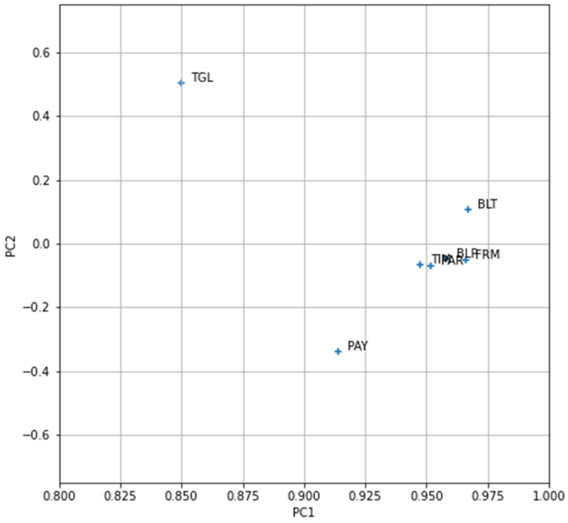

Figure 4.14. Correlations between SNOTEL time series and PC1 and PC2 time series.