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). The files we will use are in the github repository:
https://github.com/johnhorel/atmos_5040_2019
You will use the files data/gsl_yr.csv, data/utah_precip.csv, and data/utah_temp.csv and chapter_2_2019.m. For those interested in seeing the corresponding python notebook, look at chapter_2_2019.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 2018. Data are available from http://ut.water.usgs.gov/greatsaltlake/elevations/. 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. 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 necessarily representative of the state as a whole.

Code is provided to read, analyze, and plot the following figures in both Matlab and Python (chapter_2_1_2019.m and chapter_2_1_2019.ipynb, respectively). In both languages, 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. 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 ~18 in this case Hence, even though there are 122 years in the sample, only about one in eight of those values are independent of the others.

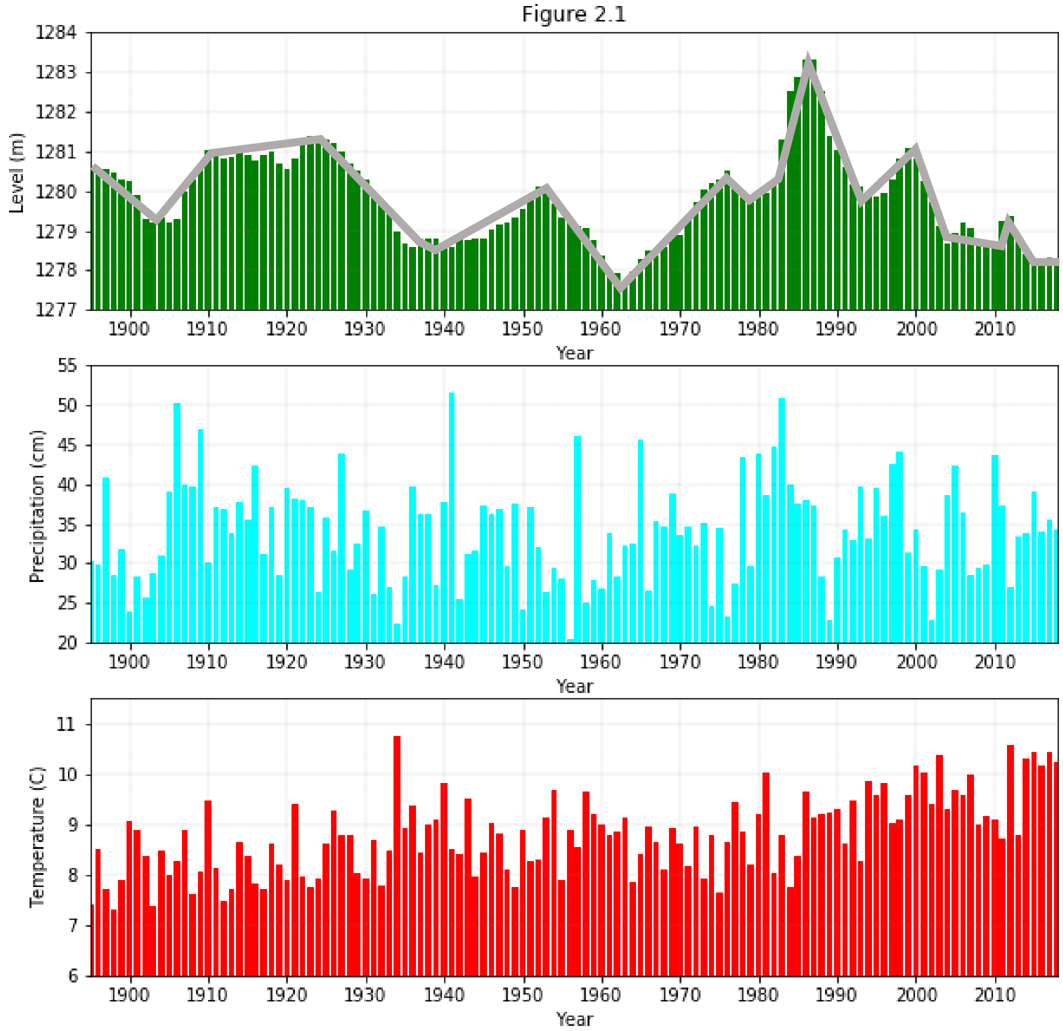

Figure 2.1. Top panel. Level (green bars, m) of the Great Salt Lake. Middle panel. Annually averaged precipitation (cm) in Utah. Bottom panel. Annually averaged temperature (ºC) in Utah.

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 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 one of the driest was 1977. 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.

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. Temperature during 1934 appears pretty unusual and is the only year where the annually averaged temperature was above 10°C before the last 1990's. Annual temperature greater than 10 °C 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, such that assuming roughly 60-70 independent values might be found in the 126 year sample. However, temperatures during the early part of the record (on or before the 1920's) certainly are lower than those of late.

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 Matlab and Python codes for that. Take a moment and look at the resulting ordered data in one of those programs. The first (last) element is the lowest (highest) value and equal to 1277.8 m (1283.3 m) for lake level. 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 data to overemphasize the total

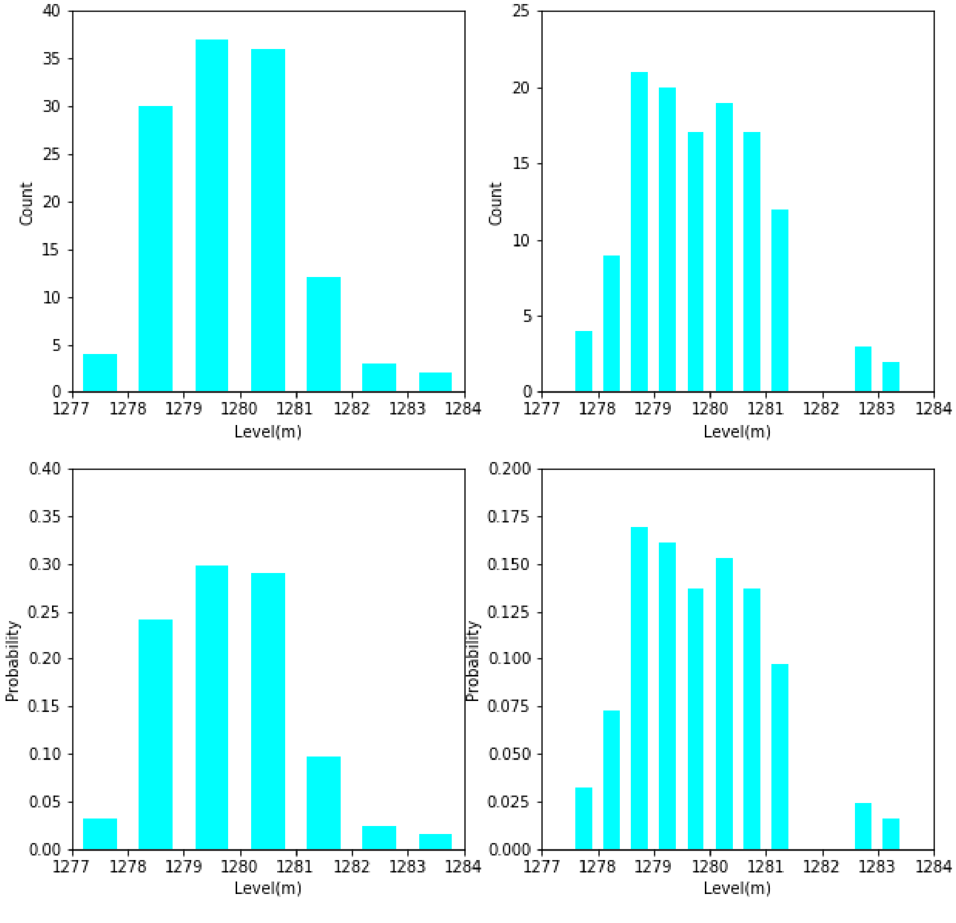

Figure 2.2. Histogram of lake level using 1m (upper left) and .5 m (upper right) intervals. Empirical probabilities that lake level falls within 1 m intervals (lower left) and within 0.5 m interval (lower right).

number of values in a sample (124 years in our case) rather than how many independent values there may be (~18).

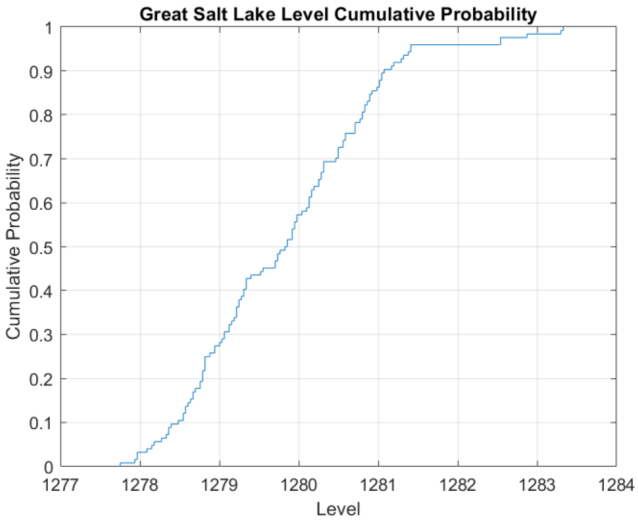

Figure 2.3. Cumulative frequency distribution of lake level.

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. So, 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 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.

The matlab statistical toolbox has a distribution fitting tool (**dfittool)** that is quite useful. We will experiment with that in class. The lower panels in Fig. 2.2 illustrate how the raw counts in each bin can be weighted by the total number of years (124) to yield empirical probabilities of how often values lie within specific ranges. For example, 70% 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). So, 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 while 30% of the time the lake level has been below about 1279 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, a probability is computed 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)

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 a peak in its distribution or whether it 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 and are easily estimated from cumulative frequency distributions or can be 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.

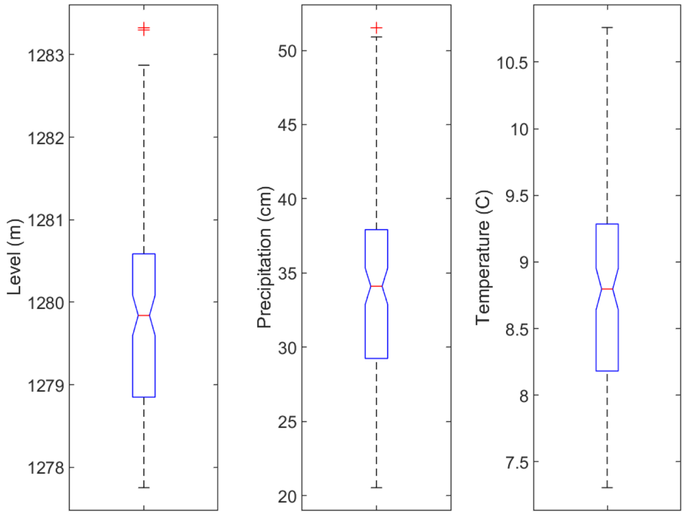

Figure 2.4. Boxplots of lake level (left), Utah precipitation (center), and Utah temperature (right).

Box and whiskers 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 'whiskers' are defined in terms of the interquartile range (IQR):

- IQR = q_{75} - q_{25}

The IQR tends to be a robust and resistant measure of spread or variation within the data set.

The top whisker is the median + 1.5 IQR and the bottom whisker is the median – 1.5 IQR. However, the Python numpy boxplot uses a different convention: the upper (lower) quartiles plus (minus) a user defined multiplier of the IQR. Finally, outliers above and below the whiskers are shown by plus symbols. In the case of the GSL level in the left panel, the high water years are viewed as outliers, which is obviously not the same as an erroneous value. It does point out that, relative to the values observed in other years, the high water years in the 80's were unusual. Similarly, the 1941 value is identified as an outlier for the annual precipitation sample in the center box and whisker plot. The Utah temperature box and whisker plot has no outliers as defined using the Matlab convention.

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.

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 variable '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.

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)*

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 difference between each value and the sample median and then taking the median of that resulting sample.

which is .91 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):

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

For the GSL level, the standard deviation is 1.13 m and the variance is 1.29 m^2. Hence, the standard deviation provides an indication that there is roughly 1 meter of variation or dispersion about the central value of 1280 m. 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:

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

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

Now we need some summation identities:

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

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

Now, if we had a very large sample, the second term $\bar{x}^2$ 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:

$\bar{x}^2 = \frac{1}{n}\overline{x^2}$ (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*.

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 negative, then data are spread more below the mean than above the mean. Skewness is neither robust nor reliant.

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

In the case of the GSL level, the skewness is .60, 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 is also positively skewed with a value of .35.

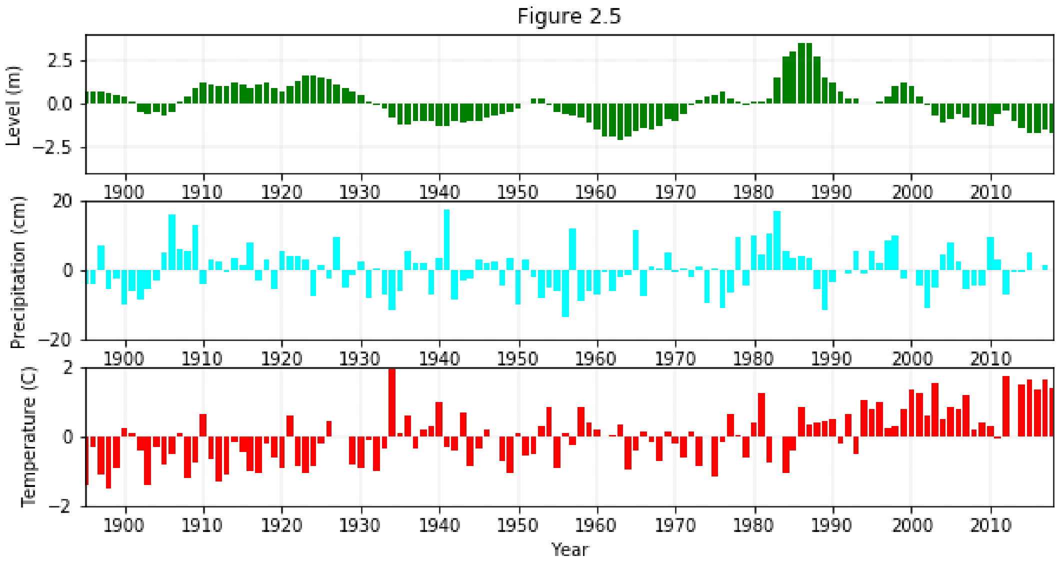

Figure 2.5. Top. Departure (m) of lake level from 124 year sample mean. Middle. Departure (cm) of Utah precipitation from sample mean. Bottom. Departure (°C) of Utah temperature from sample mean.

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′= x - $\bar{x}$ = anomaly or departure from the mean

For example, the 124-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 1981-2010 "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.

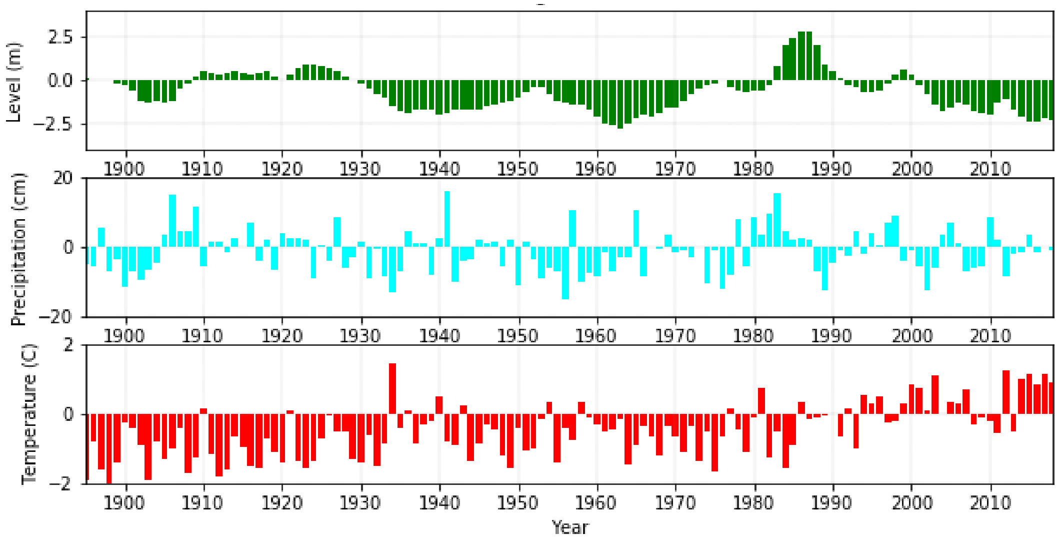

Figure 2.6. Anomalies relative to 1981-2010 climate normal.

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 warmer each month of 2018 is compared to 2017, 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. You will need the data file gsl_monthly.csv. Regular seasonal oscillations are clearly evident in Fig. 2.7, and, as shown in Fig. 2.8, the lake level peaks typically in June (after the spring runoff period) and is lowest in the fall. Using the approach that the number of independent samples can be defined by the number of line segments required, do we now have 228 degrees of freedom (one for each year's rise and fall)? NOOOO! We have 2 relevant time scales: the annual cycle as shown in Fig. 2.8 and the small number of multi-year fluctuations already described using 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. 2.7.

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" 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.,

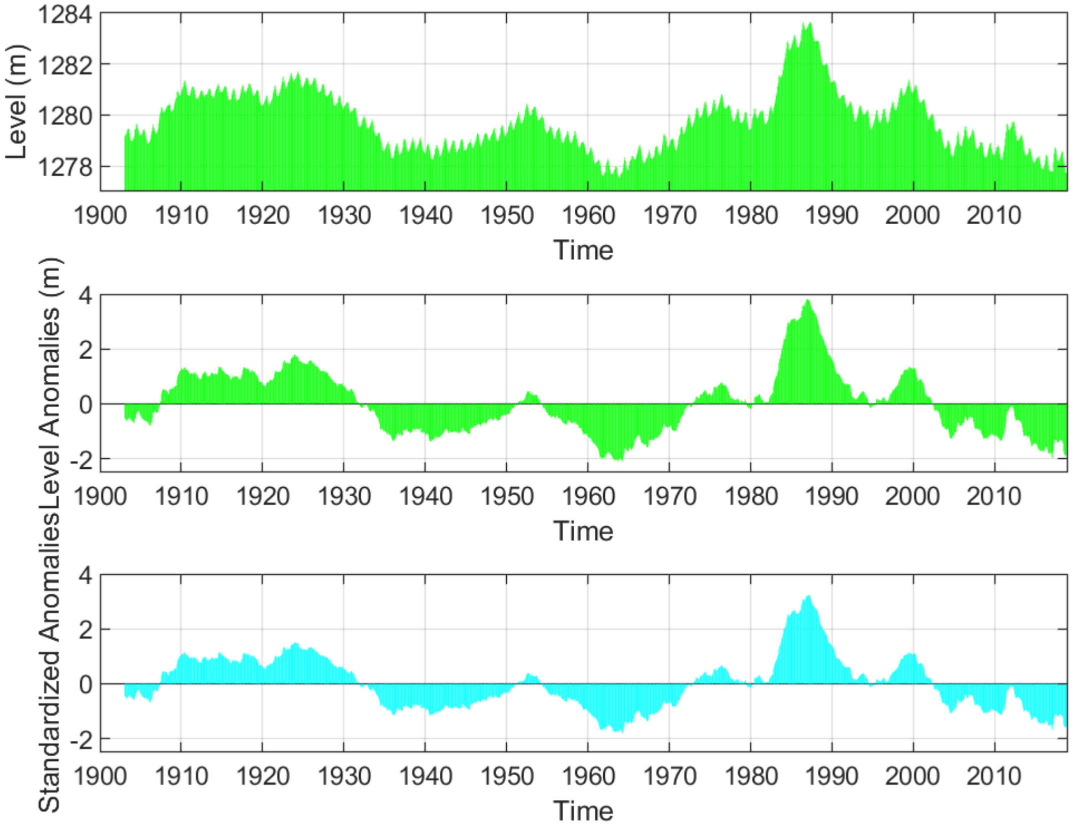

Figure 2.7. Top panel. Monthly lake level (m). Middle panel. Departures from monthly means (m). Bottom panel. Standardized anomalies (nondimensional).

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

In other words, a nondimensional time series is generated by creating 'standardized' or 'normalized' 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 1368 (114 years *12 monthly values), it is pretty clear

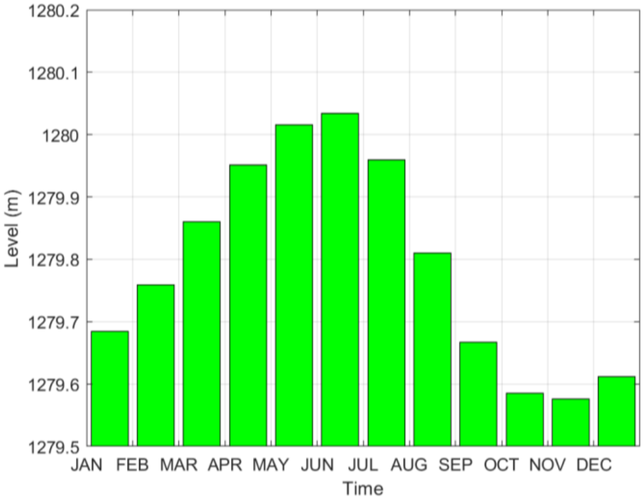

Figure 2.8. Lake level (m) averaged separately for each calendar month over the 108 years.

from Fig. 2.8 that there remain less than 20 independent samples in this time series, if we ignore some of the minor bumps and wiggles.

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.

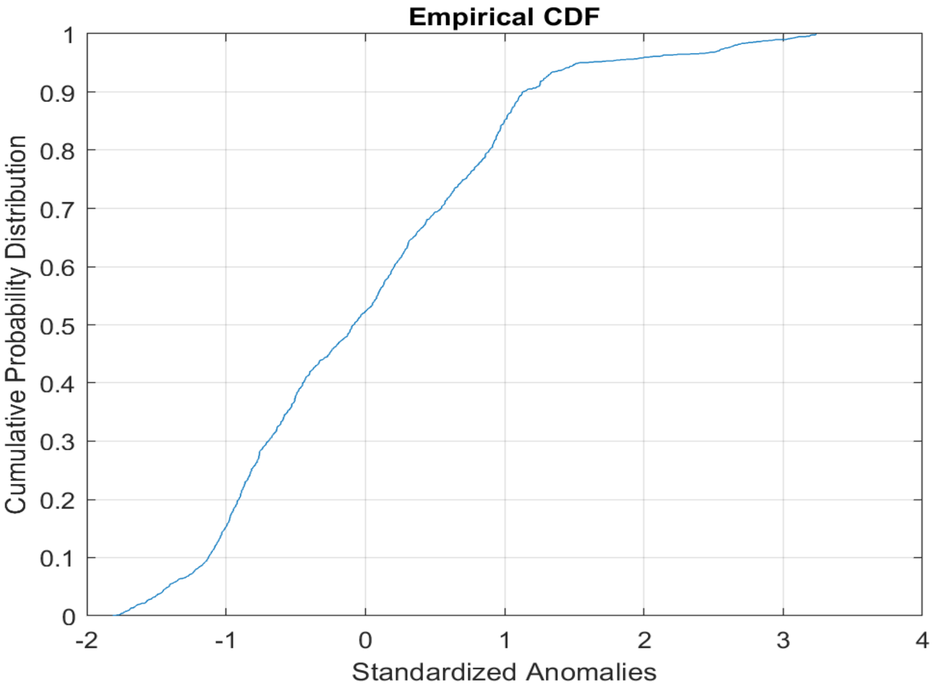

Figure 2.9. Cumulative frequency distribution for monthly values of lake level.

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 normalize. For example, let $y_{ij} = \sqrt{x_{ij}}$ and then

$$z_{ij} = \frac{y_{ij} - \overline{y_i}}{s_{yi}}$$

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”. 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).

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.

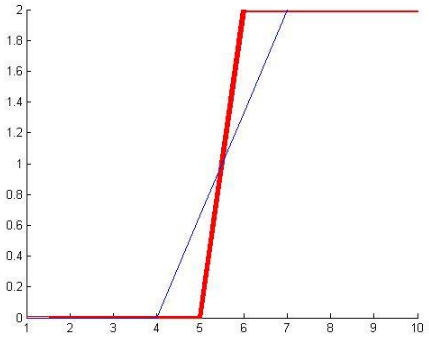

Figure 2.10. Example of original and median smoother (red line) and running mean smoother (blue line).

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 five passes with 3-year mean and median smoothers (heavy green and red lines respectively). 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. The mean and median smoothers are supplied as matlab and python functions that must be in your working directory.

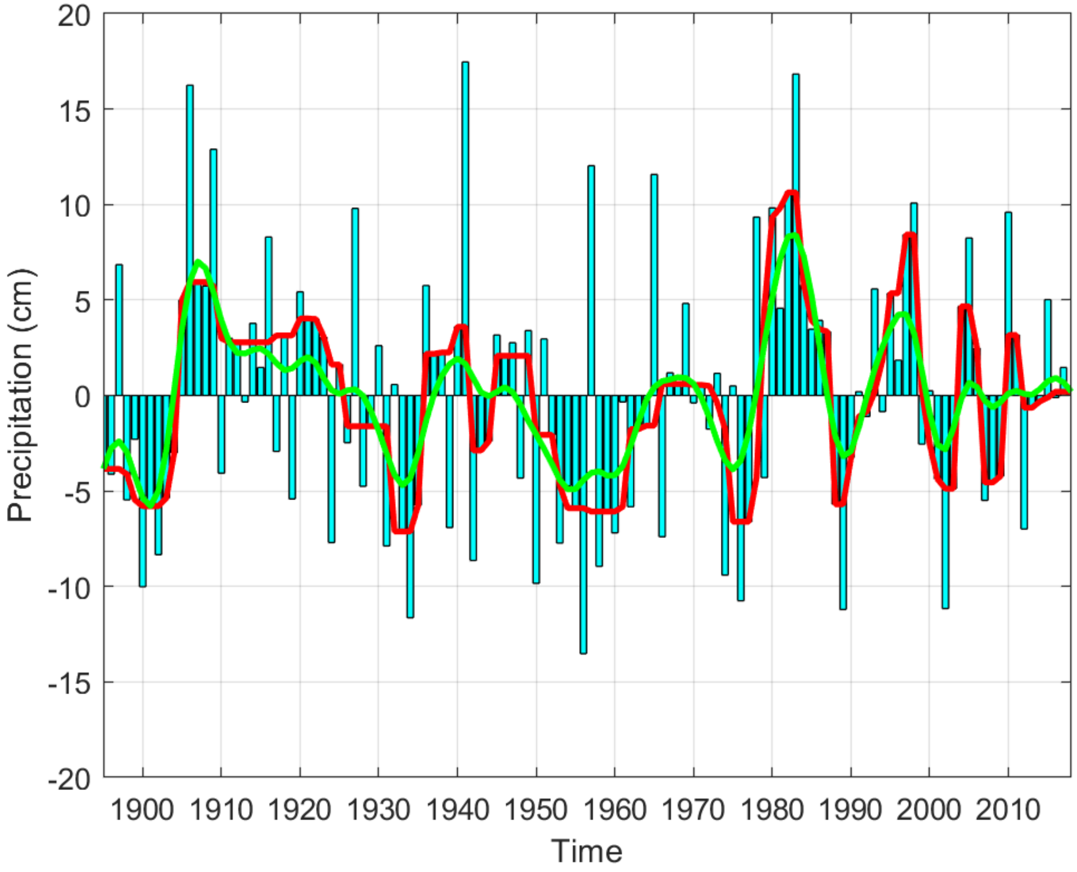

Figure 2.11. Utah annual precipitation anomalies (cyan bars) and the data smoothed by five passes of 3-point mean and median smoothers (green and red lines, respectively).

e. Exploratory Univariate Analysis of Fluid Velocity

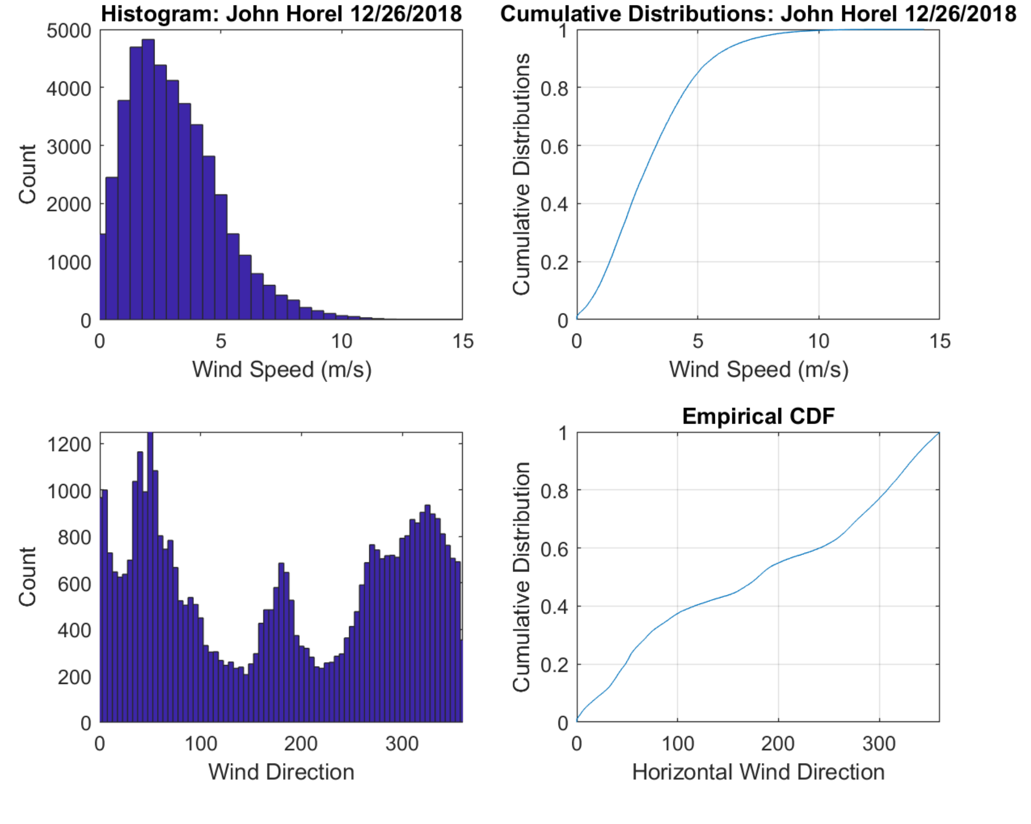

Figure 2.12. Histograms and cumulative frequency distributions of wind speed and direction.

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 $|\vec{V}|$ 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). Consider the following histograms and cumulative frequency distributions for horizontal wind speed and direction for June 2018 from the Browning Building sensor (WBB).

As shown in Fig. 2.12, the wind speeds at WBB are positively skewed with values > 5 m/s rare (only 15% of the time). Wind direction is multimodal (peaks around 40°, 180°, and 320°). There is a false peak at an angle of zero, 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. Alternatively, breaking it down into zonal and meridional components (Fig. 2.13) gives a

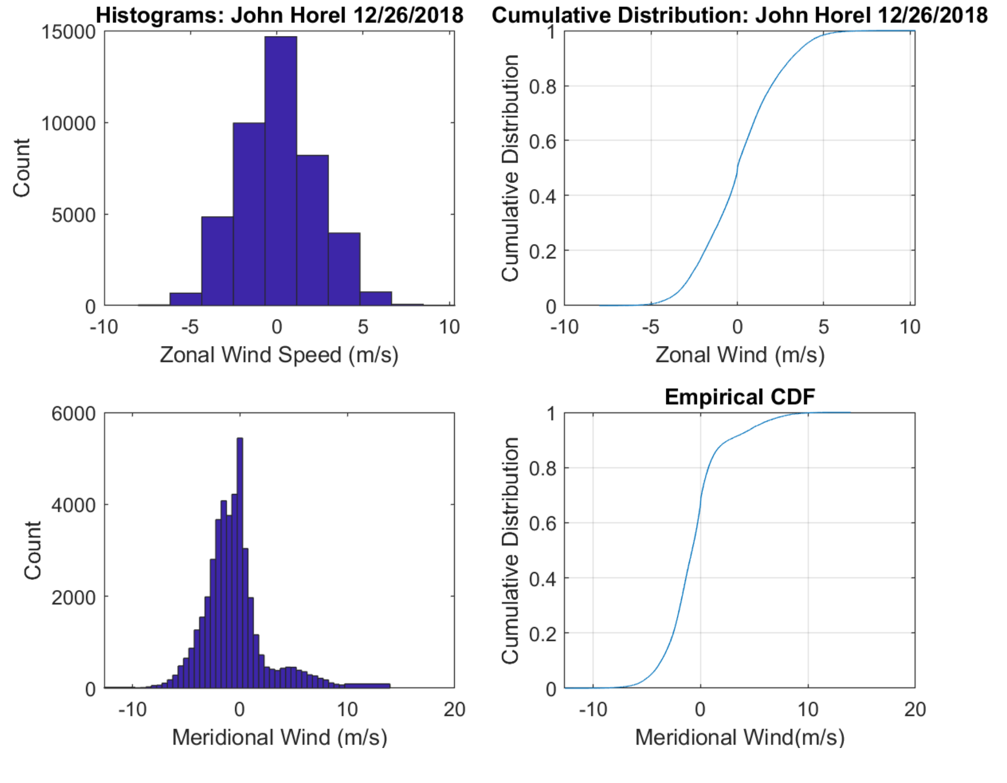

Figure 2.13. Histograms and cumulative frequency distributions of zonal and meridional wind components.

different perspective, with winds equally common from the west (positive zonal wind) and from the east (negative zonal wind) while the meridional winds are negatively skewed with winds from the south 70% of the time. 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).

As shown in Fig. 2.14, the wind rose is a special histogram that summarizes the counts from specified directions on a polar diagram. In this instance, the matlab code counts are plotted in 30^{o} increment bins for WBB and summed separately for those cases when the wind speed is less than 2 m/s (blue), less than 4 m/s (red), and all wind speeds (black). The python version is broken down into 1 m/s bins. Hence, winds are most frequent from the northeast at the Browning Bldg but also occur frequently from the northwest. The strongest winds tend to come from the northwest (greater number of counts above 4 m/s m/s).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 $\bar{V}_r = \sqrt{\bar{u}^2 + \bar{v}^2}$. Fig. 2.15 compares the mean and resultant wind speeds as a function of time of day. The strongest winds are found in the afternoon (14 MST) 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^{o} and 1^{o} should be 0^{o} not 180^{o}. The average wind

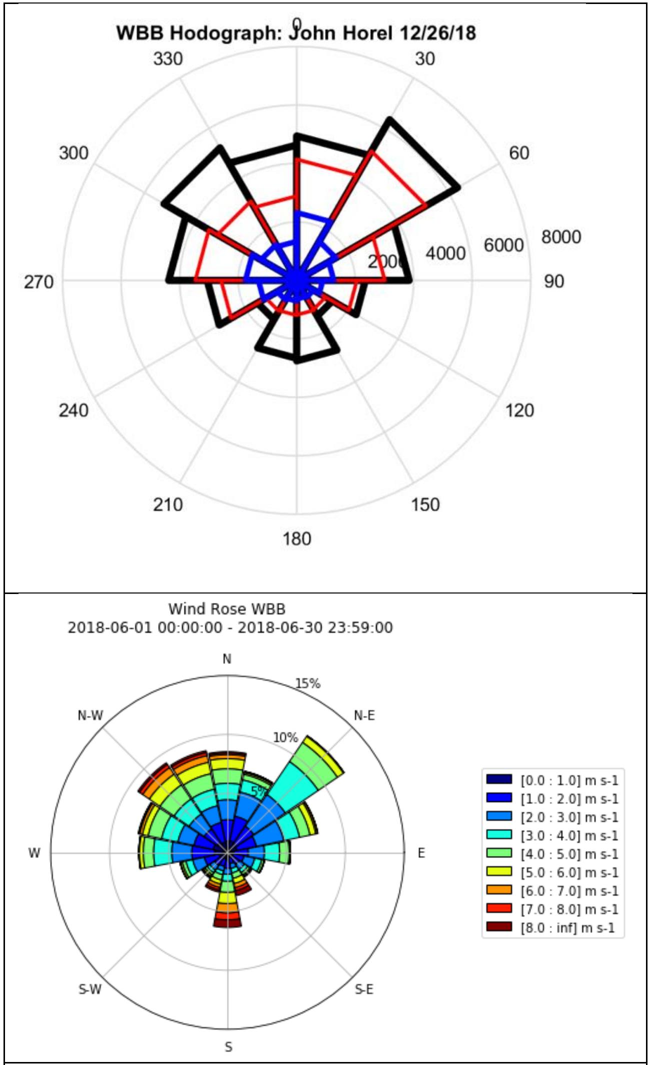

Figure 2.14. Wind rose of WBB winds during June 2018. Top panel: Matlab code. Each ring represents 500 observations of wind in 30º direction intervals and summed separately for those cases when the wind speed is less than 2 m/s (blue), less than 4 m/s (red), and for all wind speeds (black). Bottom panel: Python code. Percentage of values in 22.5º bins in 1 m/s intervals.

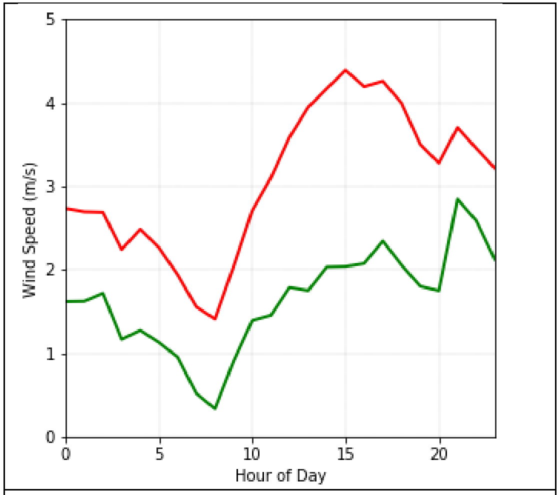

Figure 2.15. Comparison of hourly mean wind speed (red line) and resultant wind speed (green line) for each hour of the day at WBB during June 2018.

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 $\vec{V_r} = |\vec{V_r}|\hat{t}_r$where the tangential unit vector is defined from the resultant mean wind direction.