Teaching Students to Look at Their Data - a Tutorial

by William Menke
Professor of Earth and Environmental Science, Columbia University

The two most important skills that I teach to my Data Analysis students are to articulate, before even looking at their data, the properties and relationships they expect them to have, and then to critically examine them in light of these expectations. Not every expectation they have about the data will be met, of course, for new data can break down old paradigms. Nevertheless, students begin to see that data are *meaningful* - more than just the lists of numbers that they ostensibly are. The overarching goal is for the students to learn to select the data analysis methods that bring out the underlying meaning.

Data Analysis is the process of extracting useful *knowledge* from data. The knowledge usually takes the form of a physical model that connects processes operating in the world. It is abstract and often quite elegant, as well. Data, on the other hand, has a high degree of specificity and is usually terribly messy. The connection between the abstract and the specific, between elegant and inelegant, is hard to spot, especially for someone starting out in the data analysis business. Students need to be taught to focus on bringing out the connection between the model and patterns in the data.

Suppose, for example[1], that a student (or anyone else) encounters *stream flow* data for the first time. These measurements of the flow of water in rivers are an important type of environmental data and are relevant to many different issues, including agriculture, flood hazard and pollution transport. Furthermore, they are ubiquitous, with literally tens of thousands of observing stations operated worldwide. An important issue is what sequence of steps the student should be encouraged to take when first learning to work with stream flow data.

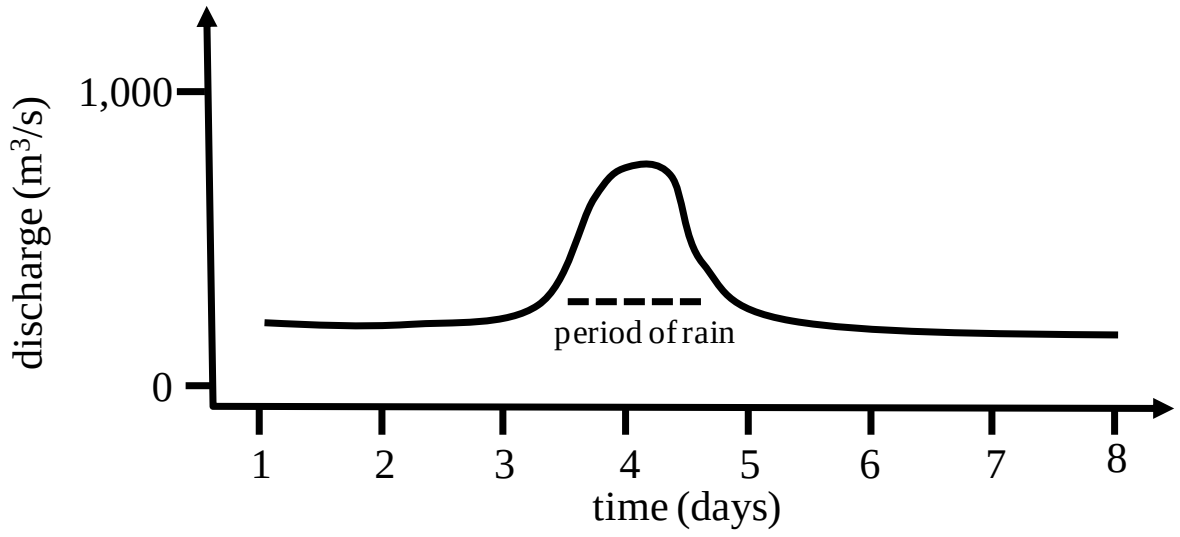

Fig. 1. "Best guess" sketch of discharge.

Stream flow is quantified by *discharge*, measured in m^3/s of water passing by a point on the riverbank. Most of my students have had sufficient experience with rivers that they can be coaxed to recollect the following facts: water flows only in the downstream direction in rivers, so that discharge should have only one sign – say, positive; the flow of water in a typical river is fairly steady over short periods of time, say minutes to hours, but can vary over longer time scales, say days to weeks; and discharge tends to increase after periods of rain. An estimate of the typical magnitude of discharge requires visualizing the movement of water in a river and relating it to volume. A large river might be 100 m wide and 10 m deep, and if the water in it was moving at 1 m/s (roughly walking speed), then a volume of $100 \times 10 = 1000$ m^3 will pass a given point along the riverbank every second – a discharge of 1000 m^3/s. I urge students to synthesize what they know about the data and make a best-guess sketch of what they expect a plot of discharge vs. time to look like (Figure 1).

The great advantage of having students visualize their expectations is that it gives them a basis for looking critically at actual data. As an example, consider the Hudson River stream flow data shown in Figure 2. Students can perform a *reality check* to verify that the data have the properties put forward in the sketch: it varies on times scales of days to weeks, as contrasted to minutes or hours; it is always a positive quantity; it has occasional peaks that might be associated with rain; and its magnitude is in accord with what might be expected for a large river such as the Hudson.

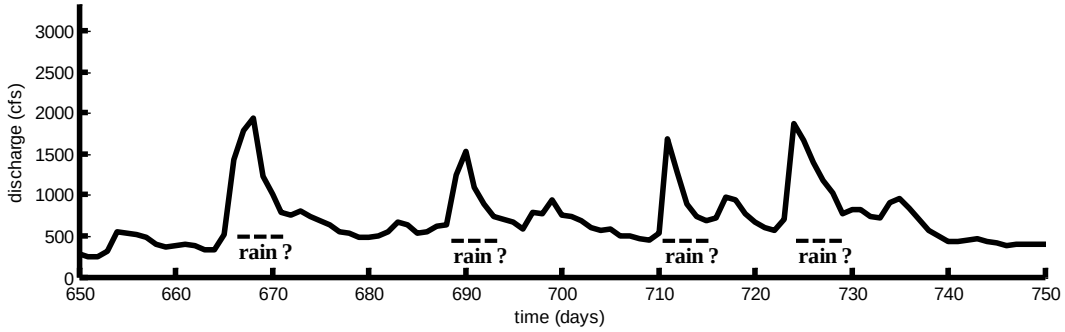

Fig. 2. Actual discharge data from the Hudson River near Albany, NY. Data courtesy of the US Geological Survey. Time scale is days after Jan. 1, 2002.

Students should be coaxed to recognize that the association of the peaks with precipitation cannot be addressed from discharge data, alone. Precipitation data is required. Obtaining such data, if it was not initially provided, requires some extra effort – but not much, since meteorological data is readily available for almost everywhere in world. A key point that I have learned from experience, and that I teach my students, is that two related datasets are almost always more easily understood when combined. In this case, taking the time to download precipitation data is very worthwhile.

Precipitation is measured in mm of water (or water equivalent, in the case of snow) that falls on a particular point on the earth. Again, students should be coaxed to jot down what they know about precipitation and to make a sketch that illustrates what they think a precipitation vs. time plot might look like. Most students have enough personal experience with weather (and weather reports) to know that: precipitation is a non-negative quantity; that the amount of rain that falls each day (at least in New York) is quite variable, with days of heavy rain intermixed with days of no rain at all, and that rainfall of an inch (25.4 mm) or so is typical of a wet day. Putting all of this together, a student can sketch a hypothetical

plot of precipitation (Figure 3B). Precipitation is best sketched together with discharge (Figure 3A), to illustrate the expected correlation between the timing of the two.

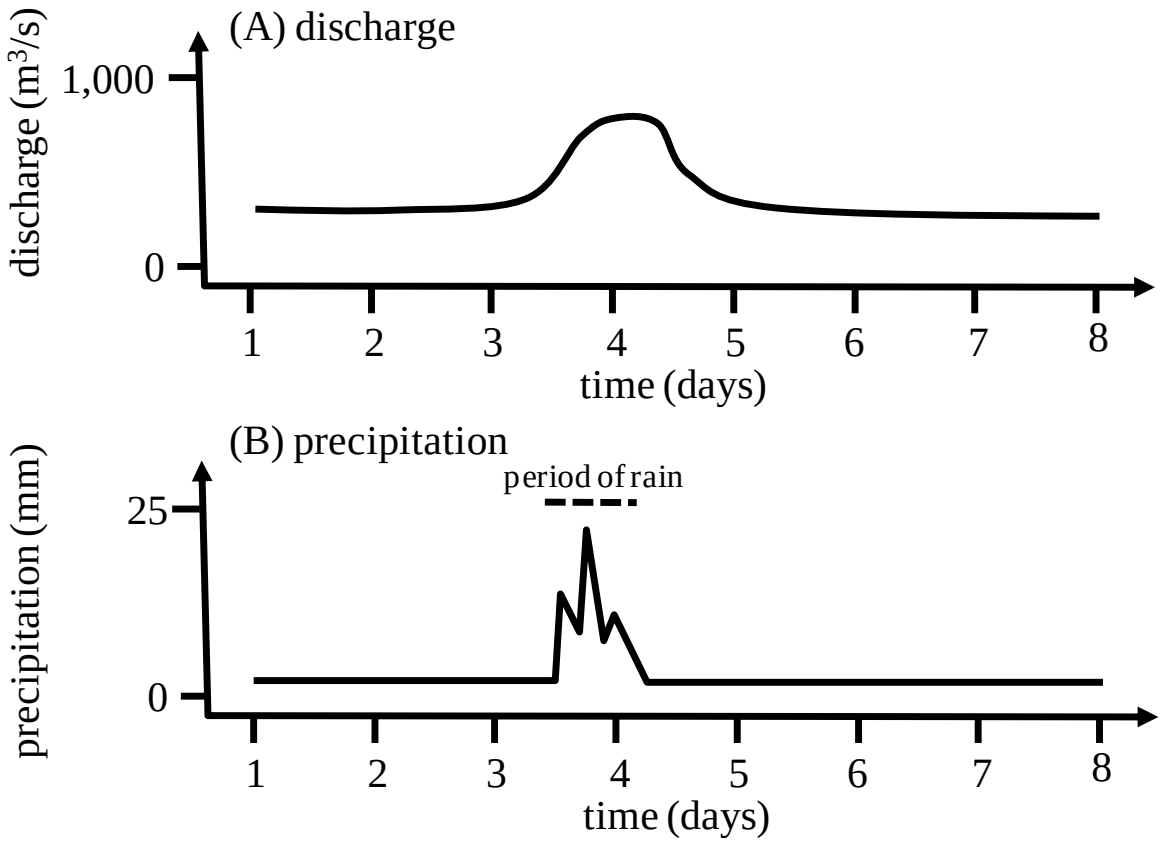

Fig. 3. "Best guess" sketch of (A) discharge and (B) precipitation.

Actual precipitation data for Albany, NY is shown in Figure 4B. Students should again perform a reality check, verifying that the plot has the properties predicted in their sketch. They might notice that while precipitation is more variable than discharge, it is arguably less variable than might be expected at times scales of hours. This is because the data are daily precipitation totals. Presumably, the rain gauge at the meteorological station is being read only once per day.

One further reality check addresses the issue of whether precipitation plausibly can provide *enough* water to account for the river's discharge. This requires using the numerical tabulation of the data, itself, and not just plots made from it. By summing up the daily total precipitation data for the year 2002, we estimate precipitation to be 0.92 meters during that year. It also requires some research on the area of the Hudson River's watershed. According the Wikipedia (2012), the river north of Albany drains about 1.8×10^{10} m^2 of land. Assuming that the Albany rain gauge is representative, the total volume of rain in 2002 is about 1.6×10^{10} m. We estimate the total volume of water flowing in the Hudson in 2002 by multiplying the discharge by 86,400 (the number of seconds in a day) and them summing over the days of the year, which gives 1.2×10^{10} m^3. Somewhat more water falls as precipitation than finds its way to the

river (the rest evaporates). The good correspondence gives us confidence that we understand the units of *both* discharge and precipitation measurements.

Students are now in the position to investigate the *model* that precipitation leads to river discharge. A cursory examination of Figure 4 indicates that, as expected, peaks in discharge correspond to peaks in precipitation (as is indicated by the vertical dashed line in the figure). However, students should be encouraged to enumerate ways in which the correspondence fails to meet their expectations. First, they might recognize that the amplitudes of the peaks do not correspond all that well. The largest precipitation occurs about day 689, but while the discharge is peaked there, three other peaks in the plot are higher. Second, they might notice that peaks in discharge are typically of longer duration than the corresponding peaks in precipitation. This *misfit* between data and prediction is very important, because it sheds light on ways that the model can be improved. Students should be encouraged to brainstorm the underlying reasons for the misfit.

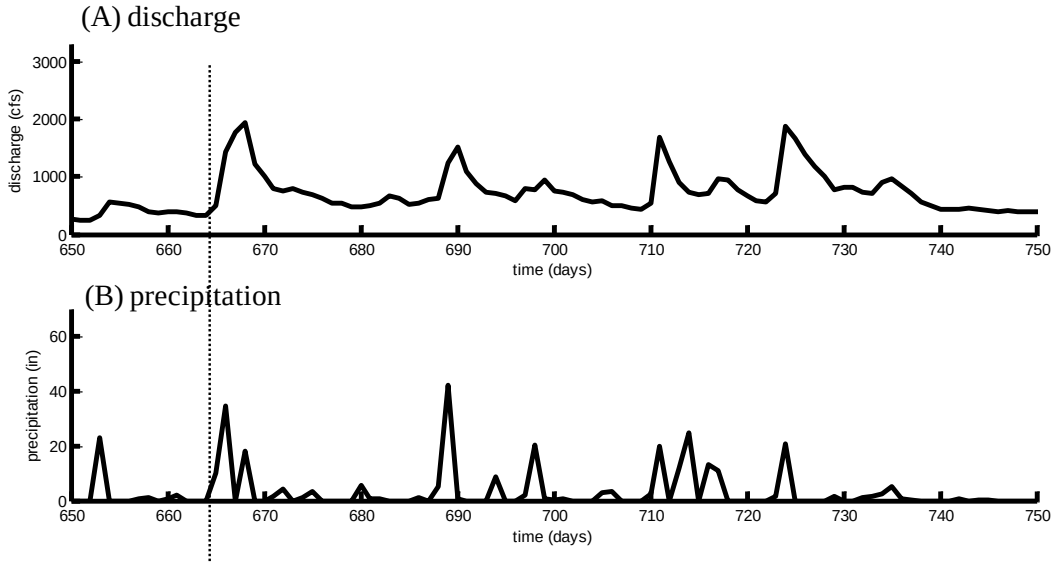

Fig. 4. Actual (A) discharge data from the Hudson River at Albany New York and (B) precipitation data in Albany, NY. Data courtesy of the U.S. Geological Survey and the National Oceanic and Atmospheric Administration. Time scale is days after Jan. 1, 2002.

Some students may realize that the amplitude problem is due to the Albany meteorological station not being completely representative of the Hudson River watershed as a whole. Albany will *over-predict* the average rainfall in the watershed if a storm hits Albany but misses other parts of the watershed. Students should be coaxed to offer ways in which the data analysis could be improved, such as by averaging several precipitation records from different parts of the watershed.

The longer duration of the discharge peaks, when compared to precipitation, sheds important light on the dynamics of the river system. Water drains from the land to the nearest stream, and then from streams to the river – a process that takes time. Thus, the watershed continues to supply water to the river well after the precipitation has stopped, though at a slowly-decreasing rate. A pulse in precipitation is *broadened* by the drainage process (Figure 5).

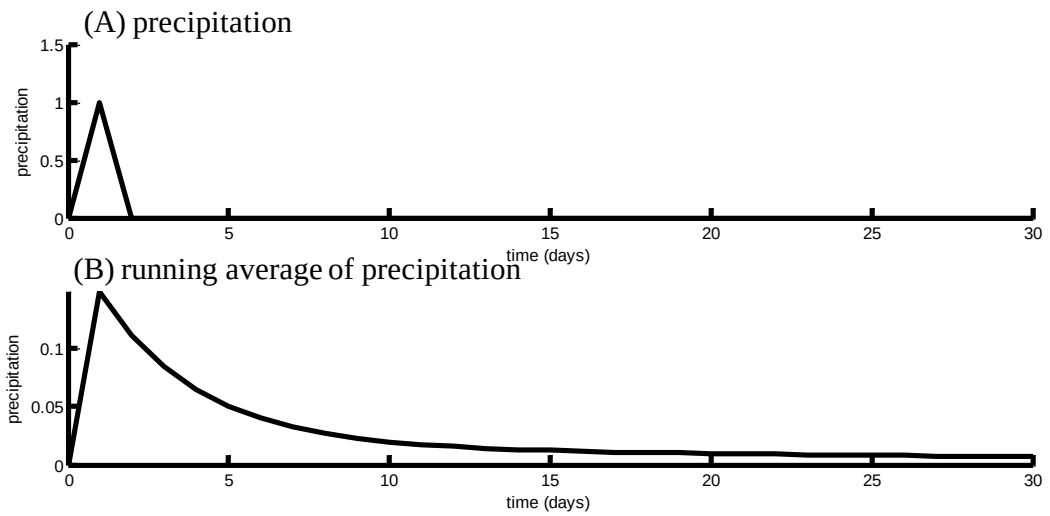

Fig. 5. (A) Hypothetical single day of high precipitation (B) Its running average, as described in the text.

Students with familiarity with advanced data analysis concepts should be encouraged to suggest ways to develop a quantitative model of the relationship between precipitation and discharge that includes the pulse-broadening process. In its simplest form, the drainage system can be thought of as delaying the arrival of different portions of the precipitation by varying amounts. Thus, if a pulse of precipitation were to be arbitrarily divided into ten equal portions, four might arrive at the observing station on the day of the precipitation, three on the subsequent day, then two and then one (for a total of ten). Thus, the discharge d_i on day i can be modeled as proportional to a weighted running average of the past and present precipitation p_j:

$$d_i = \sum_{j=1}^{\infty} w_j \, p_{j-j+1}$$

Here w_j are the weights in the running average. An interesting data analysis problem, which we will not pursue here, is to deduce the weights w_j from the discharge and precipitation data (but see Menke and Menke, 2011, Section 7.3). A similar, but simpler approach is to assume that the weights w_j decrease according to some smooth function that monotonically declines with index j, and then to select this function by trial and error. We find that

$$w_j = \exp\left(\frac{j}{3}\right) + \frac{1}{10}\exp\left(\frac{j}{30}\right)$$

produces a reasonably good match. This choice implies that the majority of the water drains away in a few days but that the last bit takes longer, on the order of months. Students can be encouraged to experiment with changing the coefficients in the formula, and to assess the effect on the results.

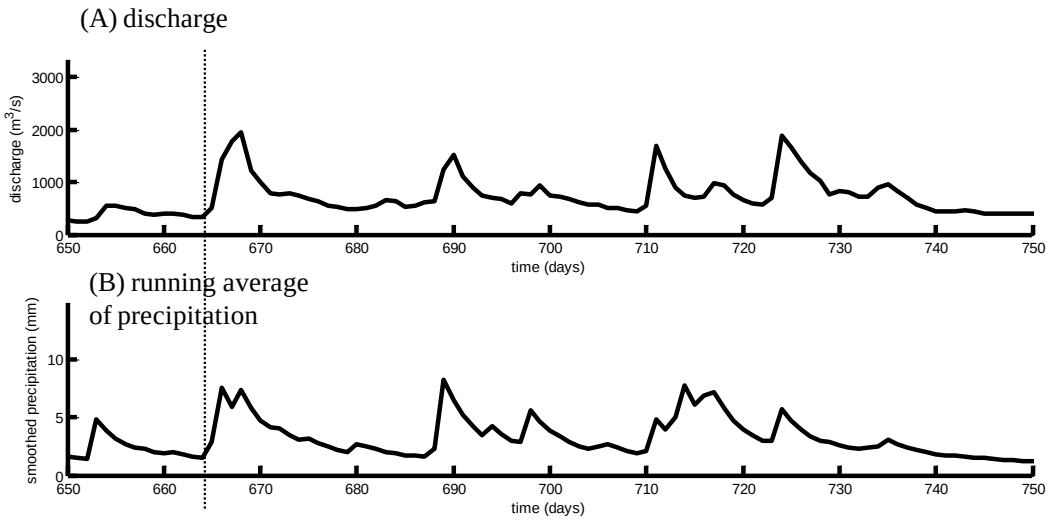

Fig. 6. Actual (A) discharge data from the Hudson River at Albany New York and (B) Running average of precipitation data, using method described in the text. Time scale is days after Jan. 1, 2002.

This systematic process of enumerating expectations about data, verifying whether actual data meets these expectations and then working to understand the source of discrepancies has several strengths. Students quickly identify and overcome misconceptions about their data and develop confidence in their ability to work with them. They learn to recognize important features and patterns. They begin to think of data, not as an opaque set of numbers, but as a tool for understanding real-world processes.

References

Menke, W. and J, Menke, Environmental Data Analysis with MatLab (textbook), Academic Press (Elsevier), 259 pp., 2011.

Wikipedia, Hudson River, en.wikipedia.org/wiki/Hudson_River, 2012.

[1]Data and MatLab scripts that implement this example are provided as a companion to this tutorial at URL http://www.ldeo.columbia.edu/users/menke/talks/datatutorial/.