UNIVERSITY OF WATERLOO

Faculty of Mathematics
Waterloo, Ontario N2L 3G1

Centre for Education in
Mathematics and Computing

Intermediate Math Circles

March 23, 2016

Data

Introduction

We live in a world in which we interact with data constantly.

On the right you see an image from satellite of the western part of Lake Erie in early spring.

There are areas of left over ice (white arrow).

And areas of an alien looking green associated with what is called the spring algae bloom.

But algae, shown in a microscope smiley face arrangement are tiny!

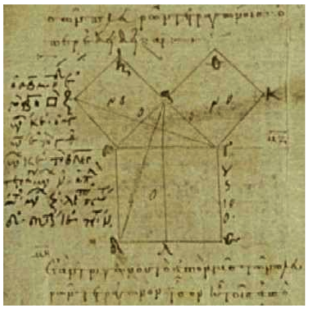
So what are we looking at in the satellite image?

Well, a satellite as a rule has to be a simple machine, usually something that detects light or some other part of the electromagnetic spectrum.

Satellite is beamed back to Earth and post-processed by a computer.

Mathematics is a very old subject (the version of the Pythagorean Theorem on the right is from the 9th century AD).

Even modern mathematics, or what you might call deductive mathematics is a few hundred years old.



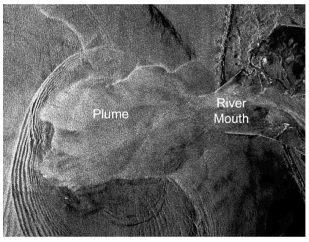
But dealing with data is a much younger, and in some sense much wilder science.

Data science or data mathematics has three components.

First you have to get the data and understand how flawed that process is.

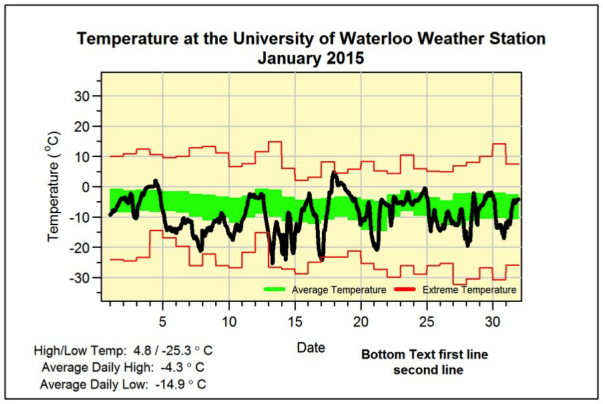
Second you have to be able to quantify the data using mathematical techniques that other scientists respect.

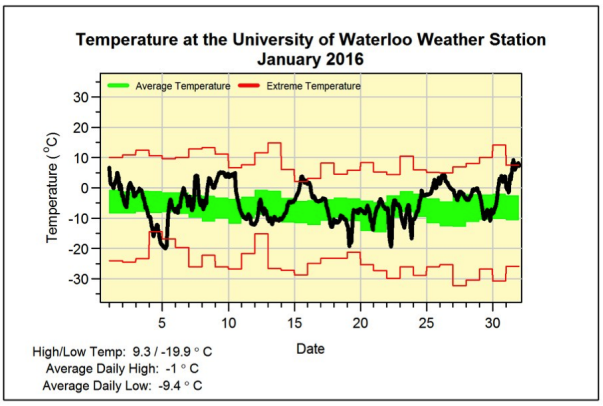
Third you need to be able to present (often in pictures) what you found.



This image is from a satellite called SAR, which uses radar instead of visible light to image a river plume

Data Example 1: The Waterloo Weather Station





This is an example of about the cleanest data you can imagine.

Air Temperature is a quantity we all relate to, and the UW weather station is a reasonably reliable source.

The raw data is plotted in black. Other quantities are plotted in red and green. Are they easy to understand?

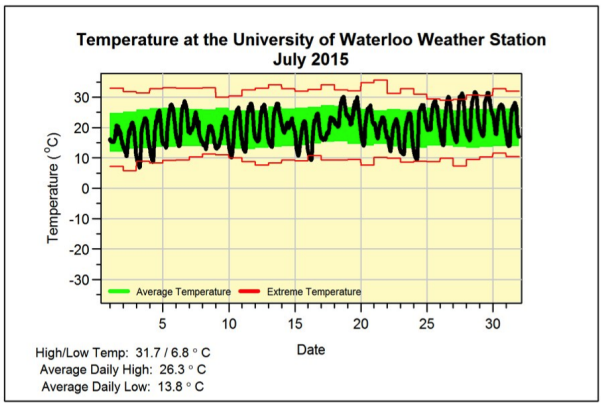
How would you compare 2015 and 2016 based on these graphs?

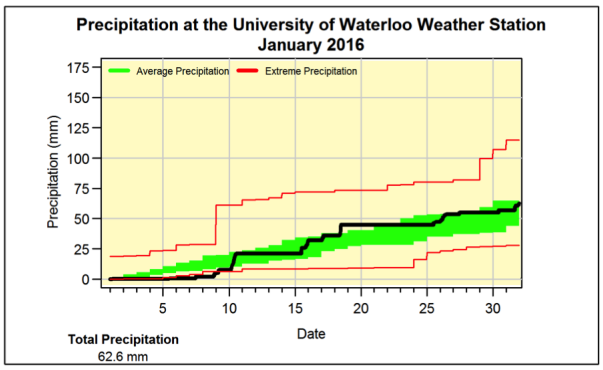
The bottom left tells you two ways, namely the average daily high and average daily low. Both were much higher this year.

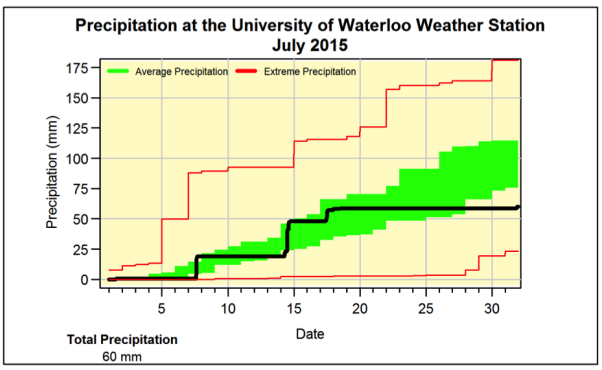
Now compare the January 2015 record and the one from August of the same year.

Obviously summer is warmer than winter, but there is a lot more to it. First of all the winter time temperature swings are larger.

Second of all the daily cycle is much more evident in the summer.







Things get a lot less tidy when we look at a variable we might care about even more, namely precipitation.

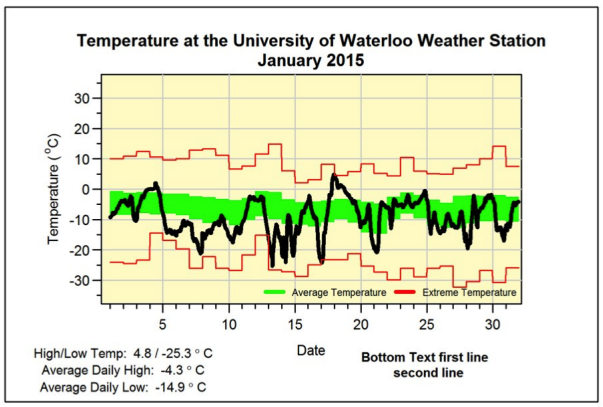
The graphs themselves are easy enough to interpret (though it is odd at first that they are presented as cumulative precipitation).

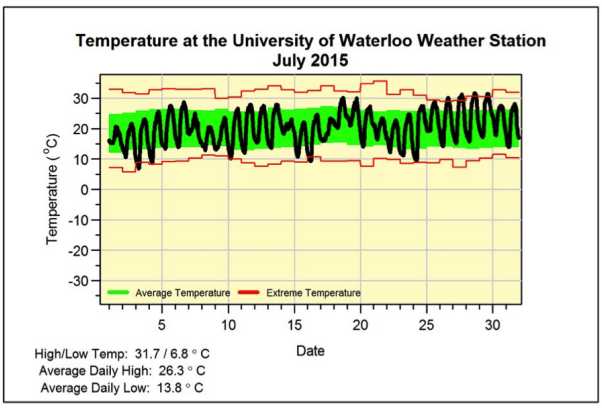
But the hardest thing to interpret is how snow and rain were compared.

Lets go back to the nice contrast result and add some actual mathematics to the story.

The independent variable here is time, and the dependent variable here is air temperature.

Notice that temperature is a numerical value by virtue of the machine used to measure it. This is imperative for us to be able to do mathematics and can be a big challenge for social scientists (for example how do we define poverty line).





The patterns are visually quite different. This is both useful and mathematically challenging since quantifying the difference is not easy (a standard technique that is well beyond the high school level is to write the data in terms of sines and cosines).

An easier to understand approach is to define quantitative descriptors of the data in the aggregate (notice all the big words; math is much simpler).

The **mean** is just the average value for all the measurements and the **variance** is the average of the difference from the mean squared.

$$< f >= \frac{1}{N}\sum_{i=1}^{N} f_i \qquad var(f) = \frac{1}{N}\sum_{i=1}^{N}(f_i - < f >)^2$$

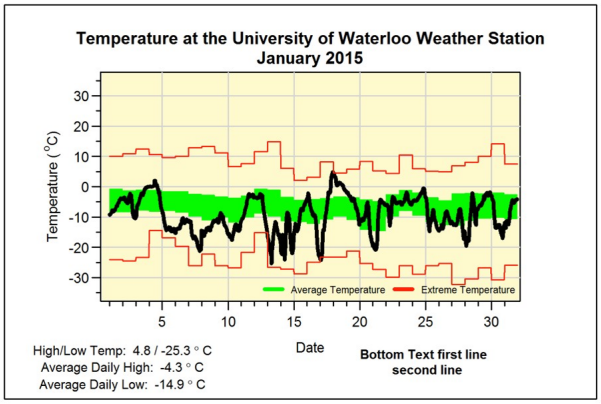
$$N = 4 \qquad < f >= \frac{1}{4}(f_1 + f_2 + f_3 + f_4)$$

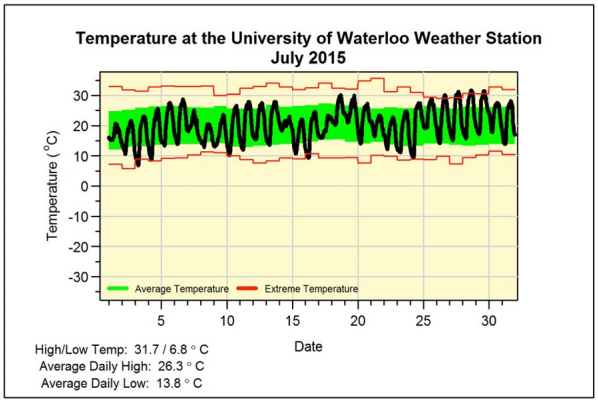
$$var(f) = (f_1 - < f >)^2 + (f_2 - < f >)^2 + (f_3 - < f >)^2 + (f_4 - < f >)^2$$

$$f_i = (5, 7, 3, 10) \qquad < f >= 6.25 \qquad var(f) = 8.9167$$

You can see right away that computing the variance is time consuming by hand, and indeed computers are much better at handling data.

- If you are paying attention you might note that the variance is very different from the mean
- This is partially due to the numbers I chose for the example, but partially due to the definition in terms of squares (incidentally, why did we square the difference from the mean?)
- Usually people report what is called the **standard deviation**, which is just the square root of the variance

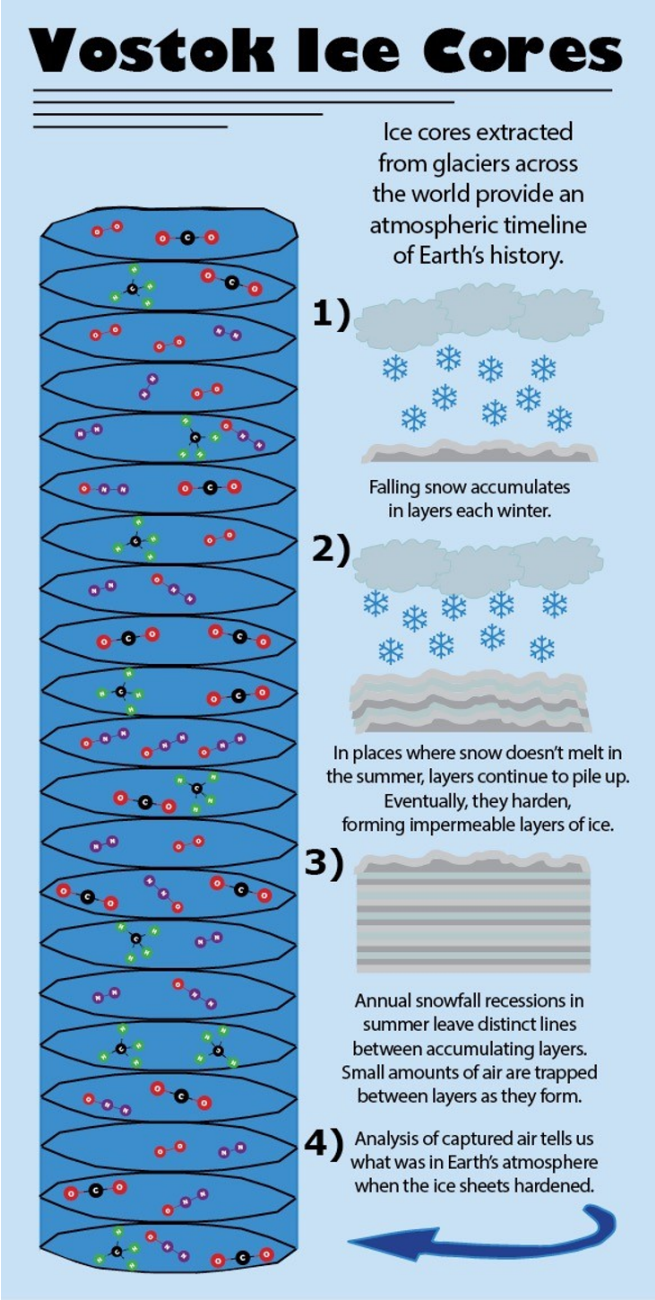




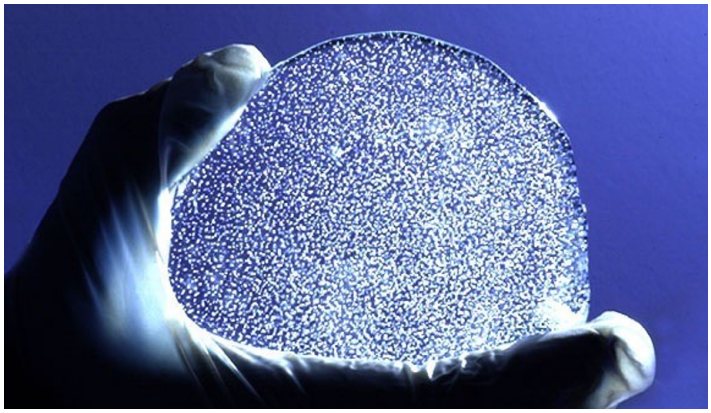
- The weather station reports the mean, which is clearly different for the two months (though how to quantify different is a more complex question)
- However the variance and standard deviation are not reported and in fact quantities that are more specific to meteorology are (the average daily high and the average daily low)
- This is an example of how context influences the choice of mathematics. Incidentally, what to do you think N is for the weather station data?

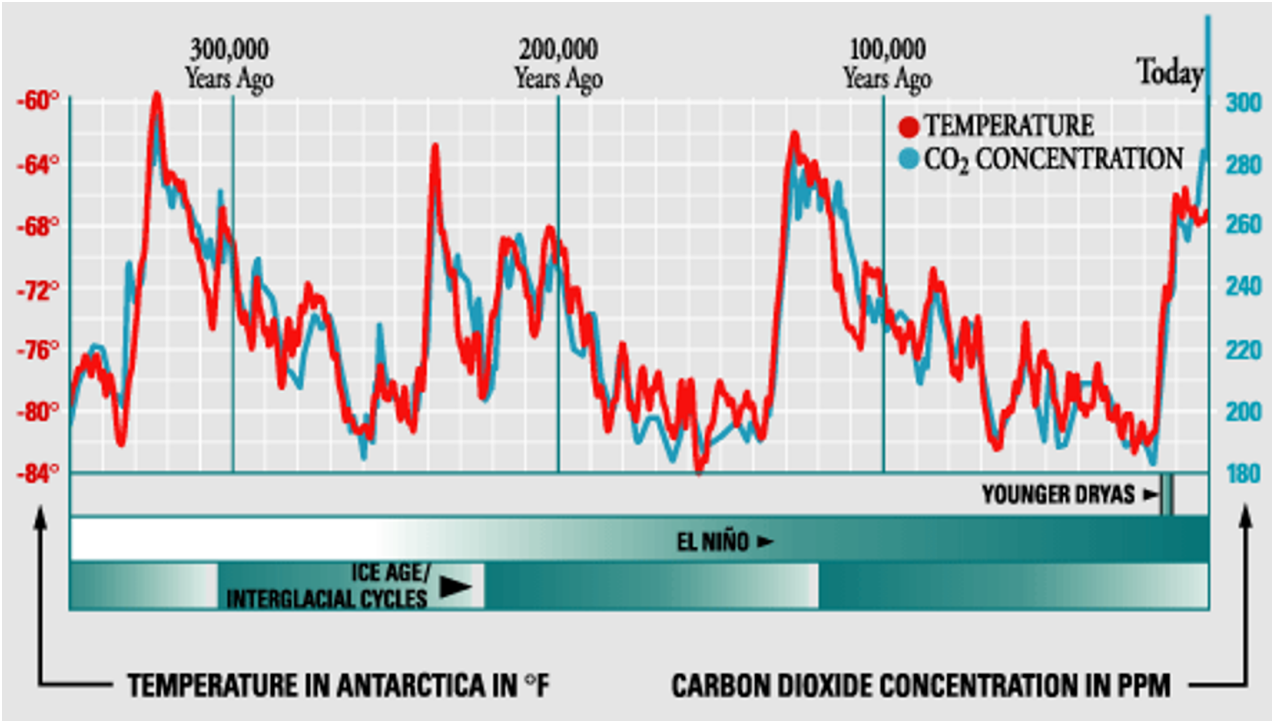
Data Example 2: The Vostok Ice Core

https://en.wikipedia.org/wiki/Ice_core

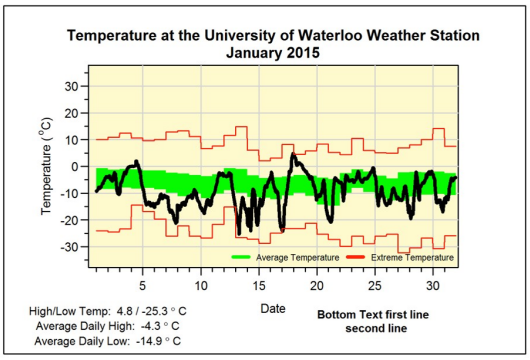


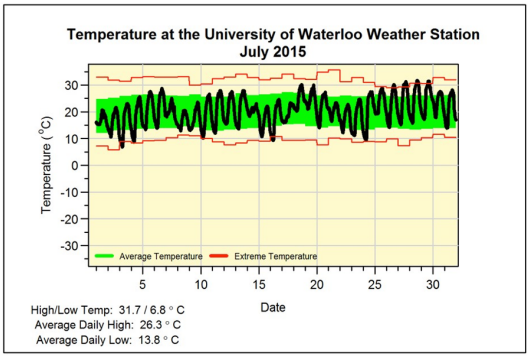
A much bigger challenge: learning about past climate from ice





With a lot of care (both scientific and mathematical) it is possible to recreate the coming and going of the ice ages from the little bubbles trapped in the ice core. This ice core is from a place called Vostok in Antarctica.

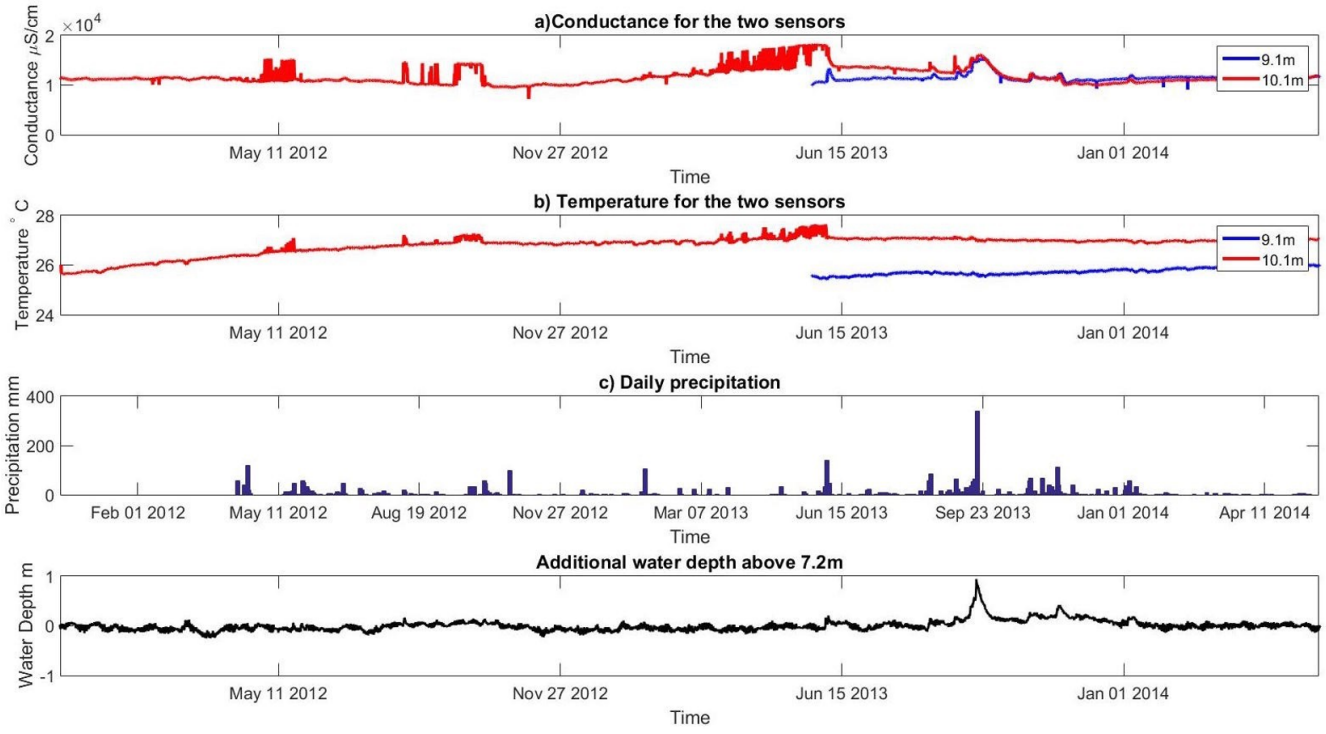


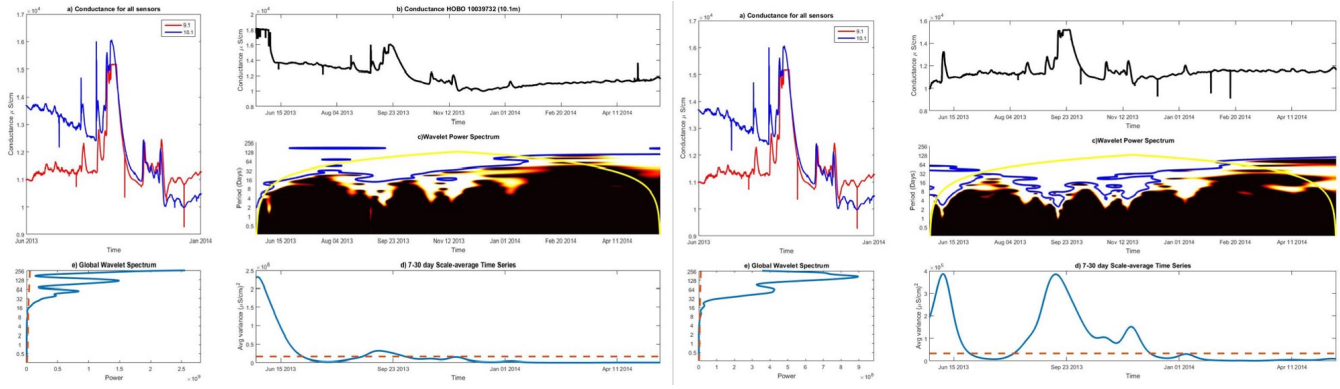


- Contrast the ice core data with the humble weather station. Both give similar looking graphs but you probably feel much more confident that you understand the weather data.
- This is because the ice core data has more actual uncertainty (is the Vostok site somehow "special") and your own knowledge of the topic is probably not as firm.
- This is part of the reason for having general mathematical definitions and techniques (these days software with a broad user group as well).
- But mathematics cannot replace individual experience.

Data Example 3:

Measurements in the Cenotes of the Yucatan





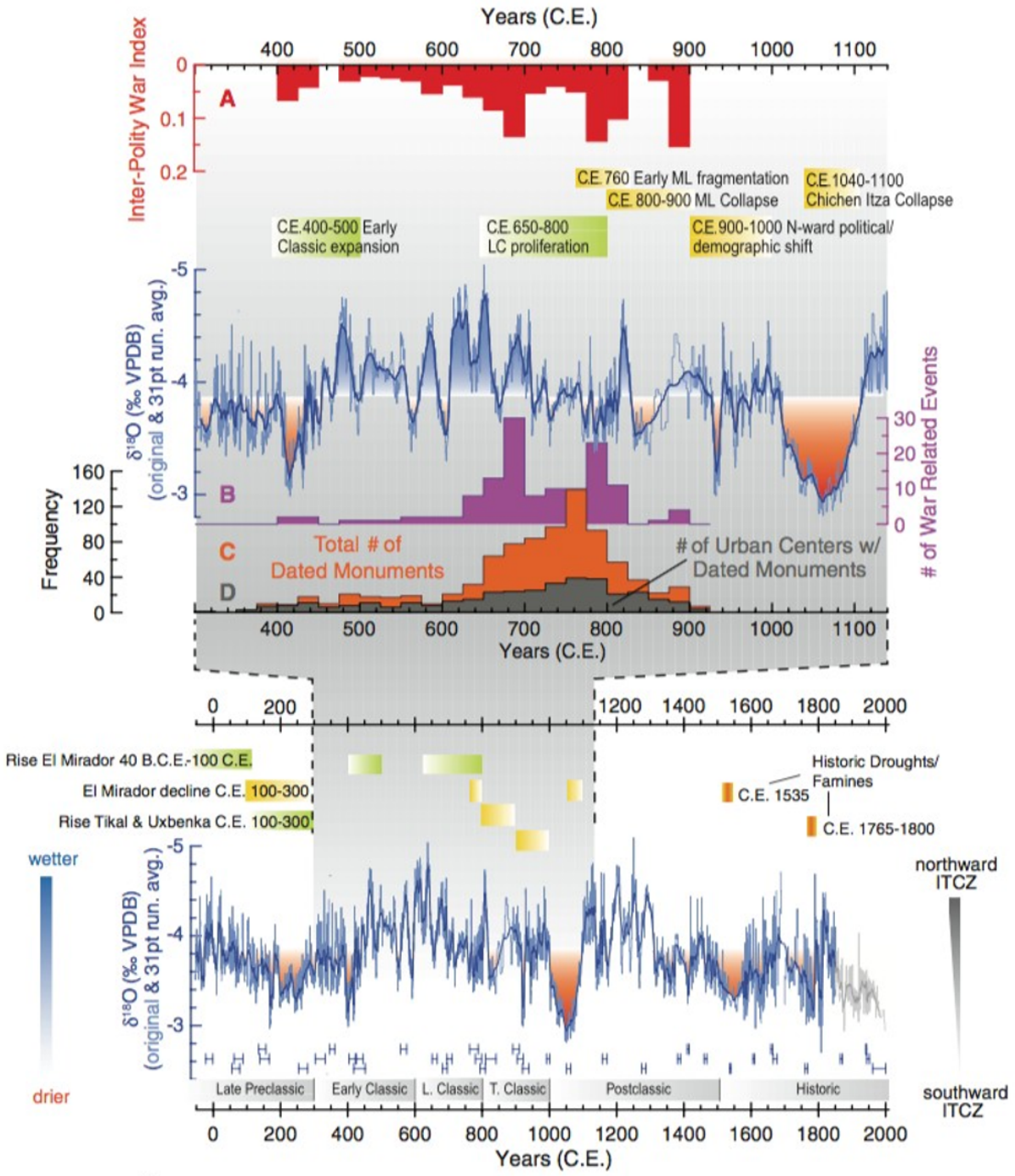


Fig. 2. (Bottom) YOK-I $\delta^{18}O$ climate record spanning the past 2000 years (40 B.C.E. to 2006 C.E.) shown relative to Maya chronology and major historical events. Blue bars just below the $\delta^{18}O$ curve indicate the small error for each of the 40 U-Th dates used to constrain the chronology of the $\delta^{18}O$ climate record (*10*). Drier-than-average conditions during this interval are shown in orange. Two historically recorded droughts in the 16th- and 18th-century C.E. accord well with the YOK-I record, and the earliest multidecadal drought in the record (200 to 300 C.E.) corresponds with decline of the large center of El Mirador and a major sociopolitical reorganization in the ML. (Top) The YOK-I $\delta^{18}O$ climate record between 300 and 1140 C.E. shown relative to major historic events along with **(A)** An interpolity warfare index based on the number of war-related events between Maya sites or rulers relative to the total number of events recorded during each interval. **(B)** Raw number of war-related events. **(C)** Frequency distribution of long-count dated monuments in the ML. **(D)** Total number of urban centers with dated monuments through time as a proxy for the development and disintegration of complex polities in the ML. All hieroglyphic data are from the Maya Hieroglyphic database (raw data is available in the supplementary materials) (*28*) and are binned in 25-year intervals. The light gray line denotes uncertainties in the 20th-century $\delta^{18}O$ record (*10*).