The Clickable Histogram (ClickHist) – A New Tool for Sifting Through Big Data in the Earth Sciences

Matthew Niznik[1] (email: mniznik@rsmas.miami.edu), Brian Mapes[1] (email: mapes@miami.edu)
[1] – Department of Atmospheric Sciences, RSMAS, University of Miami, Miami, FL

Reconciling Large Datasets and Computational Resources

For years, scientists have needed to make due with the sparse, often limiting, and ultimately disappointing availability of observations. Conversely, freely available output from state of the art climate models is now reaching unwieldy sizes, forcing the introduction of petabytes into the climate science vernacular. While traditional analysis techniques are well-suited for observational datasets of manageable size, it is not always straightforward to apply the same techniques on all available data from one or more climate models (so-called "big data"). In particular, undergraduate and graduate students may be unnecessarily excluded from such research either due to (a) having to spend years developing a sufficient programming skillset and subsequently rushing through a project at the end of their studies, or (b) limited or no access to supercomputers with sufficient memory and processing power to apply traditional analyses to large datasets.

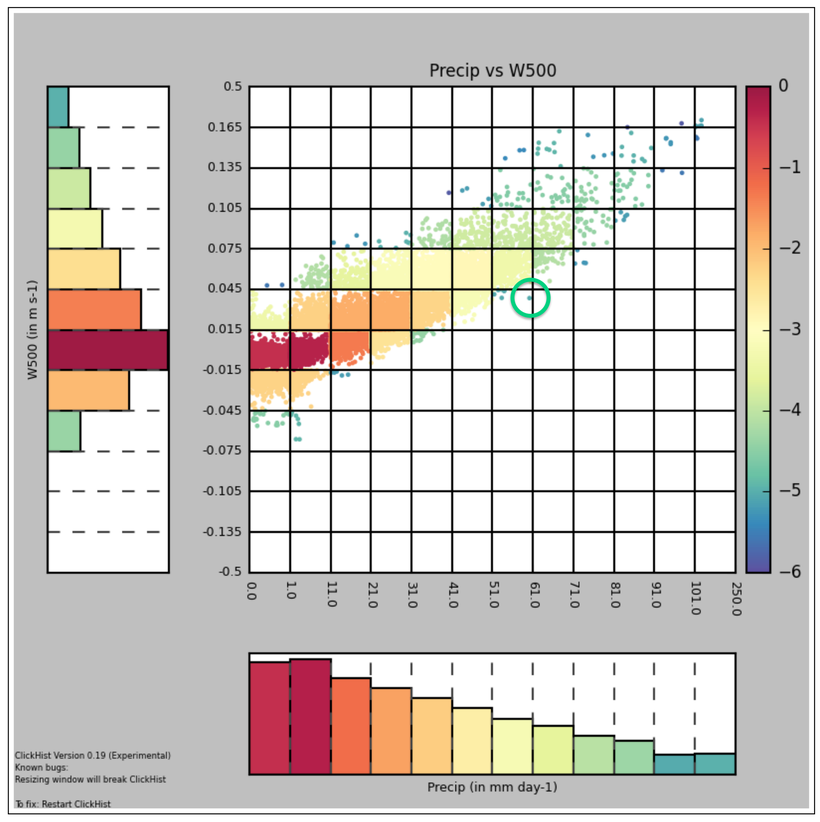

Figure 1. A sample ClickHist using the CHAD implementation for G5NR precipitation (x-axis, in mm day^{-1}) and 500 hPa vertical velocity (y-axis, in m s^{-1}). The main plot area is a combination scatter plot and 2-dimensional histogram; each scatter point is colored based on the total number of points in each bin (visually, each box). One dimensional histograms of each variable individually are also appended to the appropriate axis.

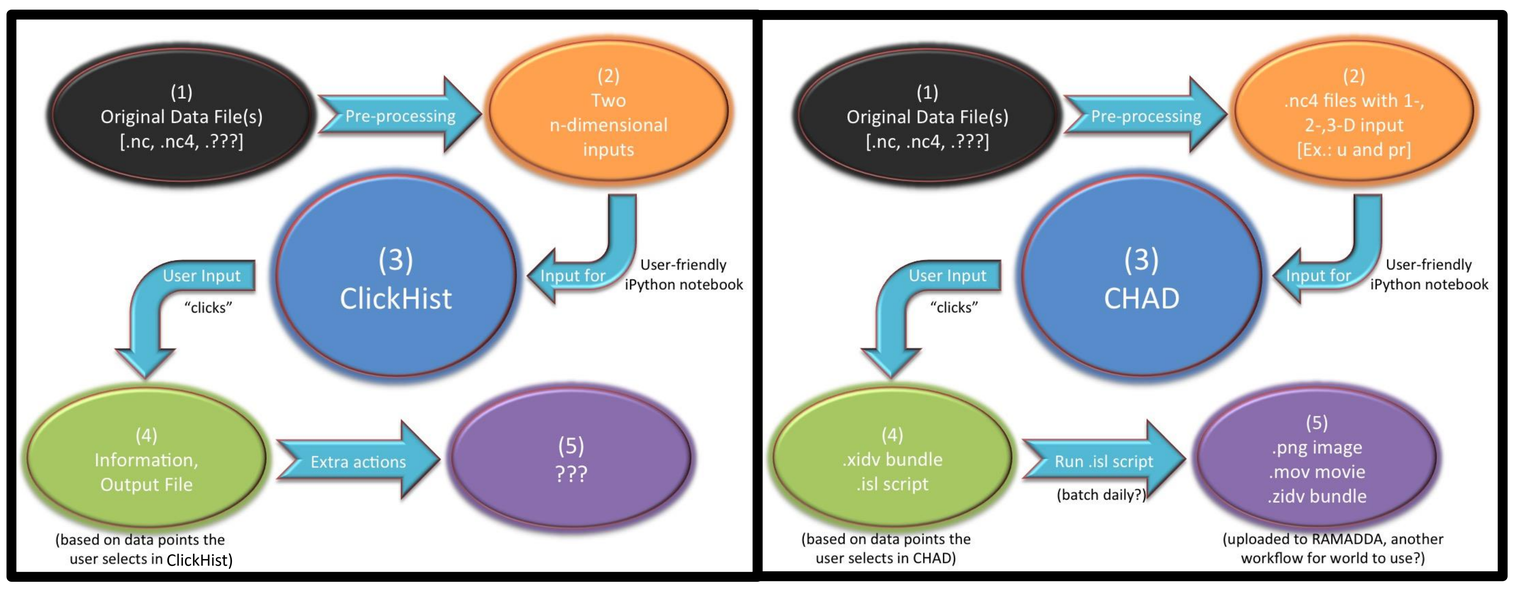

Figure 2. Flow charts showing the workflow of both the generalized ClickHist (left) and a specific application for atmospheric science, CHAD (right).

Find ClickHist on GitHub!

https://github.com/matthewniznik/ClickHist

ClickHist – One of Many Needed Solutions

Born out of one particular instance of this problem, a desire to find convective events of known rarity in GEOS-5 7-km Nature Run (G5NR), **ClickHist** addresses the current disconnect between large datasets and traditional analysis techniques. In particular, the GEOS-5 specific implementation of the ClickHist workflow is known as CHAD (ClickHist of Atmospheric Data), transforming a single scatter point into an interactive bundle input for Unidata's Integrated Data Viewer (IDV). These bundles can be shared with anyone and are perfect for case study research projects.

ClickHist is by no means limited to atmospheric science; with a little coding, the generalized ClickHist can be transformed into a powerful tool for many fields in the Earth Sciences.

Figure 1 shows a sample CHAD plotting precipitation (in mm day^{-1}) versus 500 hPa vertical velocity (in m s^{-1}) for two years of the G5NR in the region 160°W–120°W, 20°S–20°N, while **Figure 2** shows the workflow design of both ClickHist and CHAD. **Figure 3** shows a still image of a sample case study bundle in IDV generated from the CHAD in **Figure 1**.

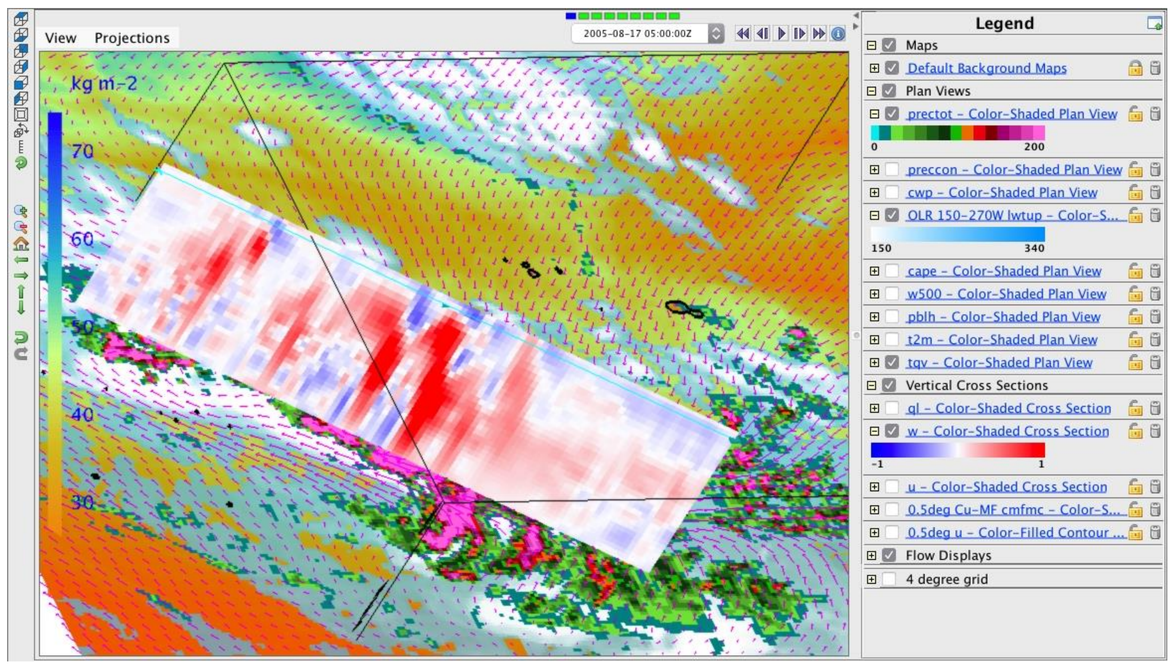

Figure 3. A sample case study bundle in IDV associated with the circled point in Figure 1, centered on 152°W, 20°S. Shown are precipitation (greens and pinks, in mm day^{-1}), outgoing longwave radiation (grays and whites, in W m^{-2}) surface temperature (warm colors, in K), 800 hPa wind vectors (pink, in m s^{-1}), and a cross-section of vertical velocity (reds and blues, in m s^{-1}). In IDV, the bundle can be interactively rotated and animated; plots of fields (the five shown or the many others not shown) can be turned on and off and combined as desired.

We are eager to collaborate with scientists and students across many fields – please get in touch if you're interested!

Thanks to Bryan Raney, Brian Matilla, and Marie McCrary for testing ClickHist along the way.