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

Matthew Niznik1 (email: mniznik@rsmas.miami.edu), Brian Mapes1 (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.

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.

Find ClickHist on GitHub!
https://github.com/matthewniznik/ClickHist

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.