



Research | October 01, 2018

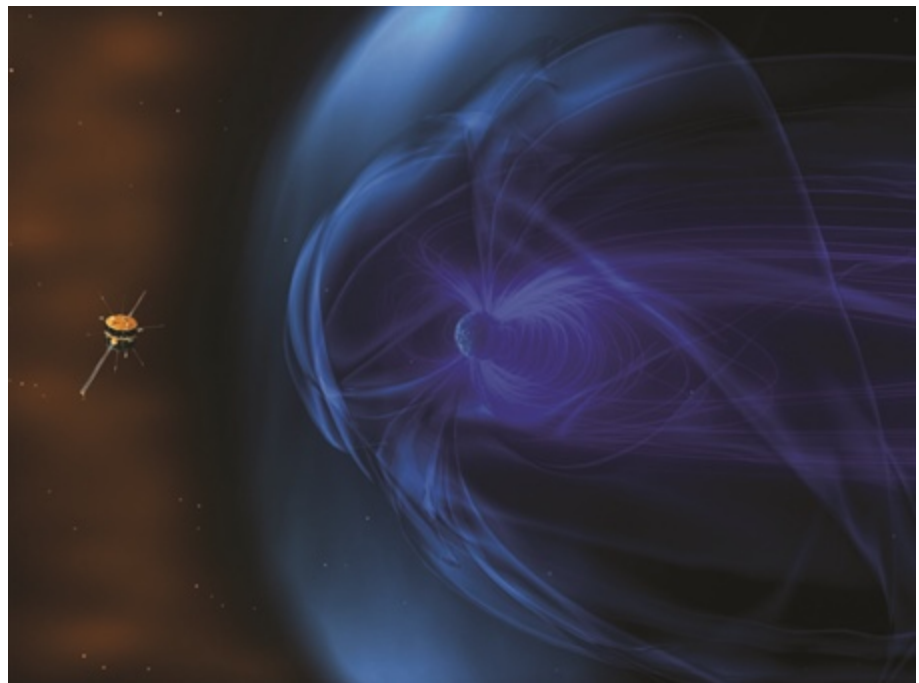
Information Theory in Earth and Space Science

By Joshua Garland and Elizabeth Bradley

Recent advances in information theory—coupled with vast improvements in the extent and resolution of Earth- and space-science-related time series data—can help answer some of the biggest questions facing humanity. Such developments enable our understanding of the dynamics of abrupt climate changes and allow us to assess the predictability of geomagnetic storms.

Until fairly recently, time series data sets with enough accuracy, length, and temporal resolution to support information-theoretic analysis have been hard to come by in some areas of Earth and space science. Replicability is also a major issue. Drilling a three kilometer core through an ice sheet and analyzing each centimeter in a spectrograph is an expensive proposition, as is launching a spacecraft to sample conditions on Pluto. However, recent developments in laboratory techniques have improved ice core sampling resolution by an order of magnitude, and the number of satellites observing the solar system has increased remarkably. Other long, high-resolution geoscience data sets are also available. These advances provide a host of exciting opportunities for the applied mathematics community to make meaningful contributions to the fields of Earth and space science using information theory.

For example, in 2005, Tom March, Sandra Chapman, and Richard Dendy applied mutual information—a measure of how much one random variable reveals about another—to observations made by NASA's Wind spacecraft to trace solar-wind effects at different points on Earth (see Figure 1) [6]. Scientists have also used other members of the family of entropy



measures nucleated by Claude Shannon's work in Earth and space

Figure 1. The Wind spacecraft from NASA's Heliophysics System Observatory has spent nearly two decades observing particles and measuring crucial properties of the solar wind before it impacts Earth's magnetic field. Image courtesy of NASA.

sciences. However, some associated challenges persist. For instance, calculation of the Shannon entropy rate from a real-valued time series requires symbolization¹ of the data, a procedure that is fragile in the face of noise and susceptible to biased results. *Permutation entropy* [1] sidesteps these issues with ordinal analysis—conversion of the sequence of real values into value-ordered permutations—to symbolize the data. Researchers have used both binned Shannon entropy and permutation entropy to explore the predictability of different climate change events captured in El Niño–Southern Oscillation proxy records derived from Laguna Pallcacocha sedimentary data [8].

Permutation entropy techniques are particularly useful in the study of ice core data. For example, the 3,300-meter West Antarctic Ice Sheet (WAIS) Divide core captures climate samples from the past 68,000 years. Figure 2 shows the weighted permutation entropy (WPE)—a variant of a technique that uses a normalization to de-emphasize noise effects [3]—calculated in sliding 500-year windows across a water-isotope trace from that record. The large jump in WPE between five to eight kiloyears (ka) was initially puzzling. Lab records solved the mystery, showing that an older instrument—used to analyze that segment of the core—introduced noise into the data, an effect that was not visually apparent in the δD data itself. Outliers that are all but invisible in the raw data also leave clear signatures in the WPE, in the form of square waves as wide as the calculation window (this is visible at roughly 17, 26, and 47 ka in Figure 2). In addition to detection of data problems, information-theoretic analysis can also lead to fascinating *scientific* knowledge. For instance, the patterns in WPE values reveal possible signatures of geothermal heating at the core's base and appear to correlate closely with accumulation [5]. Another observation of interest is the absence of any signature of the large, abrupt *Dansgaard-Oeschger* events—which punctuated the last glacial period—in the WPE trace [2], suggesting that these events may not represent significant changes in the climate system's information mechanics [4, 5].

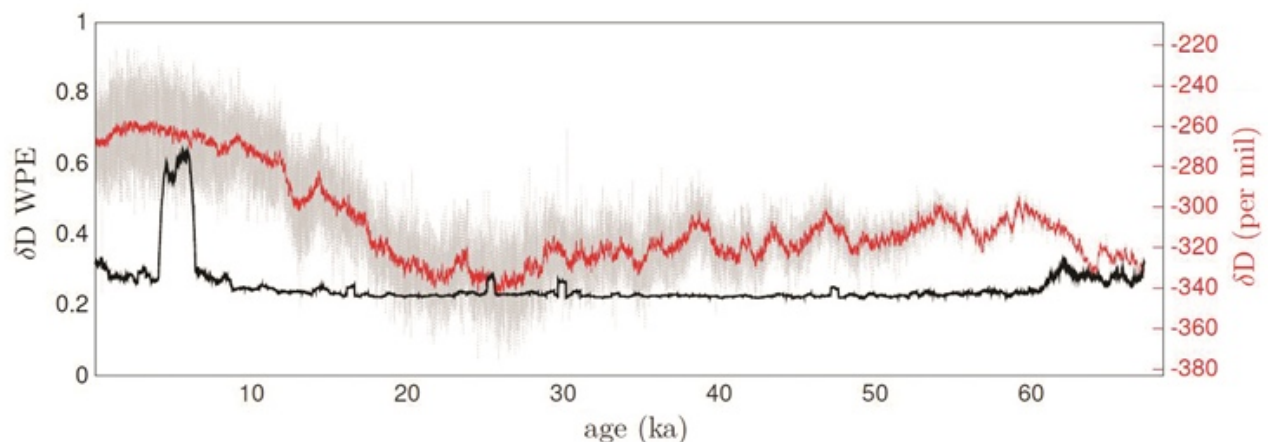
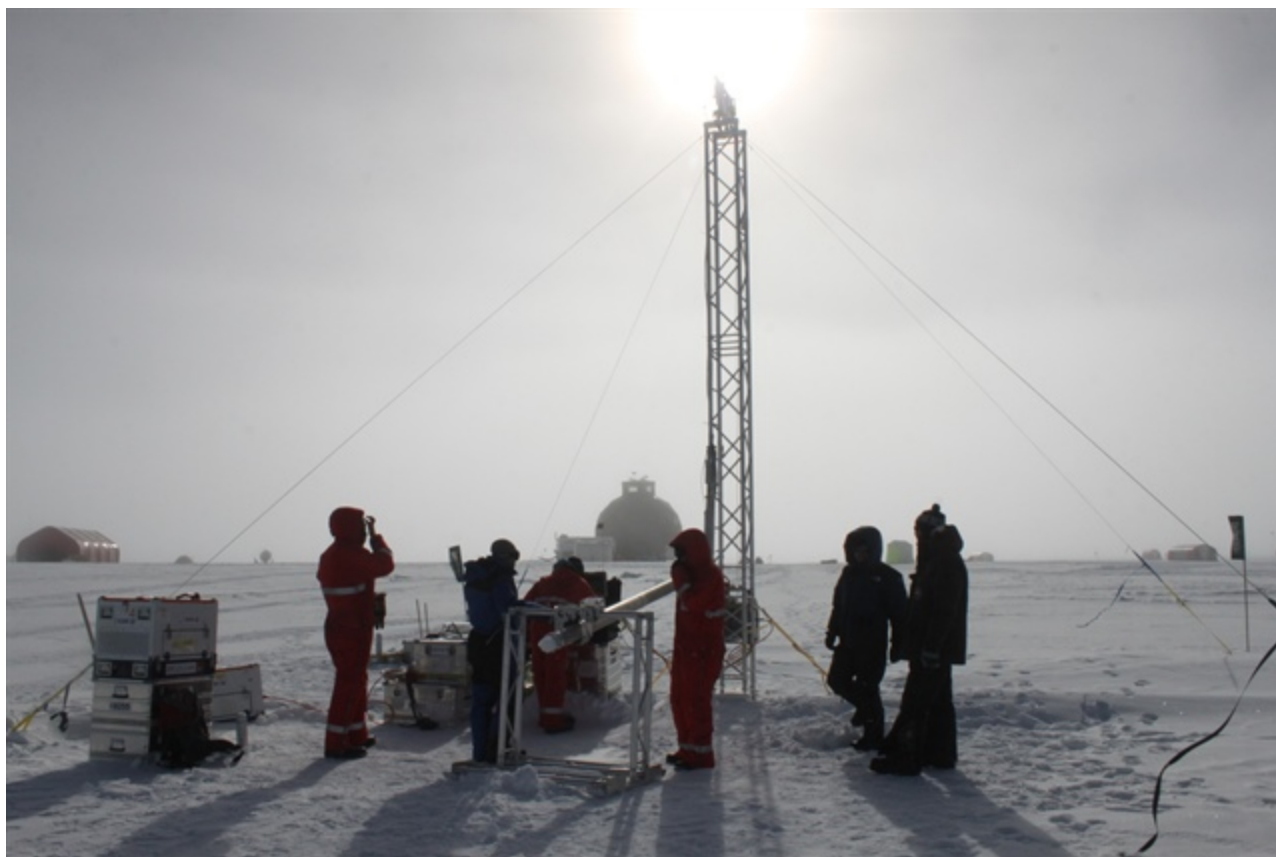


Figure 2. The deuterium/hydrogen ratio (δD) measured from the West Antarctic Ice Sheet Divide core in kiloyears (ka) before present day. The original data is shown in grey, the smoothed data (500-year moving average) is in red, and the

weighted permutation entropy (WPE) calculated from the original data is in black. WPE values run from 0 (no new information and fully predictable) to 1 (all new information and completely unpredictable). Figure courtesy of [4].

Timelines from Earth and space science data sets are often irregular, and observations tend to aggregate in strange ways. For example, ice cores are sampled at evenly-spaced intervals in *depth*, but spaced nonlinearly (and unevenly) in time because the core's upper layers compress the lower layers. Indeed, multiple factors affect these timelines; thickness of ice core layers depends on annual accumulation rate, diffusion mixes isotope data through the ice over time and space, and some sections of data go missing altogether between ice sheet and laboratory. Sensor data rates vary wildly during long space missions like New Horizons due to hardware issues, power allocation choices, and cloud occlusions that impact Earth-observing satellites.

Such timeline irregularities pose a problem for *any* rate-based calculation, information-theoretic or otherwise. To work around them, even out timelines, and fill in gaps, scientists use methods ranging from linear interpolation to complex physics models. These strategies can have profound—and heretofore unexplored—effects on the signals' information content. Understanding these ramifications is an interesting mathematical problem, as they can cause spurious short-term correlations and regularities that skew outcomes. For instance, the “ramps” introduced by linear interpolation create artificial permutations that are repetitive and completely predictable, which lowers the permutation entropy. Since interpolation plays an increasingly large role as one delves deeper into an ice core, this effect is depth-dependent. Timelines in sediment cores, where material can be carried away by currents and bioturbated by marine organisms, may also be problematic. Interpolation is used routinely in these situations as well, generally without consideration of its repercussions on the data. The mathematics necessary to understand this kind of data preprocessing's impact on information content is still under development, although researchers have made some recent progress in the general area of irregularly-sampled data [7, 9].



Climate researchers on the site of the Eastern Greenland Ice-core Project use a drill to collect ice core samples. Public domain image.

The scientifically critical problem of significance testing is also an issue when working with data sets like the water isotope record from the WAIS Divide or the solar-wind temperature on Pluto, which are expensive to gather and all but impossible to replicate. Significance testing or uncertainty quantification with only a single data set is nearly impossible. But this is also changing; the new South Pole Ice Core provides replicate data in a few segments, a number of new ice core drilling projects are underway around the world, and recent technological advances can vastly improve time series data pertaining to Earth and the solar system.

A key question about any event is whether it is an expected, natural part of the associated system—e.g., changes in seasonal solar insolation at a point on Earth—or unexpected and random, like the impact of a large asteroid or a coronal mass ejection. Information theory has the power to answer these and other important Earth and space science questions, and the resolution and extent of time series data are improving to the point of supporting these analyses. These developments are inspiring heightened interest in this area from both the geoscience and applied mathematics communities.

¹ Conversion of numerical exam scores to letter grades using a set of bins.

References

- [1] Bandt, C., & Pompe, B. (2002). Permutation entropy: A natural complexity measure for time series. *Phys. Rev. Lett.*, *88*(17), 174102.
- [2] Dansgaard, W., Johnsen, S.J., Clausen, H.B., Dahl-Jensen, D., Gundestrup, N.S., Hammer, C.U., & Bond, G. (1993). Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature*, *364*(6434), 218-220.
- [3] Fadlallah, B., Chen, B., Keil, A., & Príncipe, J. (2013). Weighted-permutation entropy: A complexity measure for time series incorporating amplitude information. *Phys. Rev. E*, *87*(2), 022911.
- [4] Garland, J., Jones, T.R., Bradley, E., James, R.G., & White, J.W.C. (2016). A first step toward quantifying the climate's information production over the last 68,000 years. In *International Symposium on Intelligent Data Analysis 2016. Advances in Intelligent Data Analysis XV* (pp. 343-355). Stockholm, Sweden.
- [5] Garland, J., Jones, T.R., Bradley, E., Neuder, M., & White, J.W.C. (2018). Climate entropy production recorded in a deep antarctic ice core. Preprint, *arXiv:1806.10936*.
- [6] March, T.K., Chapman, S.C., & Dendy, R.O. (2005). Mutual information between geomagnetic indices and the solar wind as seen by wind: Implications for propagation time estimates. *Geophys. Res. Lett.*, *32*(4).
- [7] McCullough, M., Sakellariou, K., Stemler, T., & Small, M. (2016). Counting forbidden patterns in irregularly sampled time series. I. the effects of undersampling, random depletion, and timing jitter. *Chaos: Inter. Jour. Nonlin. Sci.*, *26*(12), 123103.
- [8] Saco, P.M., Carpi, L.C., Figliola, A., Serrano, E., & Rosso, O.A. (2010). Entropy analysis of the dynamics of El Niño/Southern Oscillation during the Holocene. *Phys. A: Stat. Mech. Appl.*, *389*(21), 5022-5027.
- [9] Sakellariou, K., McCullough, M., Stemler, T., & Small, M. (2016). Counting forbidden patterns in irregularly sampled time series. ii. reliability in the presence of highly irregular sampling. *Chaos: Inter. J. Nonlin. Sci.*, *26*(12), 123104.

Joshua Garland received an M.S. in applied mathematics and a Ph.D. in computer science from the University of Colorado Boulder. He is currently an Omidyar Postdoctoral Fellow at the Santa Fe Institute. Elizabeth Bradley holds an S.B., S.M., and Ph.D. from the Massachusetts Institute of Technology and has been a member of the University of Colorado's Department of Computer Science since 1993. Her research interests include nonlinear dynamics, time series analysis, and artificial intelligence.