Measuring what matters in the era of big data

Easton R White
Department of Biological Sciences, University of New Hampshire, Durham, NH

Over the past few decades, we've witnessed an explosion in the amount of data available to ecologists. We can now measure the greenness of the planet from satellites; track the movements of individual organisms across the globe; and obtain real-time, high-frequency information from sensor networks distributed across land, air, and aquatic systems. But is the current interest in big data distracting us from measuring what truly matters?

Clearly, so much ecological research involves careful experimental design and considerations of statistical power. But not every hypothesis can be tested with experiments. Here, I am more focused on observational studies with large, often publicly available, datasets. Much of my own research has concentrated on this type of work. Monitoring for the sake of monitoring is important as it can lead to surprising results or new questions we never envisioned. At the same time, I believe that, at both individual and institutional levels, we need to be thoughtful about how we design new monitoring programs or use data from existing programs.

In some cases, the right variables often prove difficult to measure, while the wrong ones remain within easy reach. For example, imagine you are studying what may be driving invertebrate population dynamics in a temperate estuary. Temperature loggers cost little to deploy, and temperature data may already be available from existing monitoring programs. Each logger can collect millions of datapoints over a short time window, even if there is little variation over time. In addition, we may have equal rationale to consider other variables, such as dissolved oxygen or pH, which are harder and costlier to monitor. The sheer volume of temperature data and relative ease in its collection can create the illusion of importance, but convenience is not the same as relevance. What's more, when our response variables and predictors are constrained by what data are available, the scope of questions we can ask is also limited.

The same dynamic plays out with new technologies—from eDNA to acoustic recorders to GPS tags—that generate reams of new data. These tools expand what we can measure, but they don't tell us what we *should* measure. Too often, our technological tunnel vision drives the questions we ask, drawing attention away from the data that may be harder to collect but ultimately more important.

The abundance of data also brings new challenges. With large datasets, issues of data quality and bias can easily go unnoticed, creating a false sense of confidence that more data automatically translates into better science. In hypothesis testing, very large samples reduce standard errors, making even trivial relationships appear statistically significant—though they may have little or no biological meaning.

To overcome these challenges, we need to return to the roots of our discipline. What questions do we want to address? By choosing the questions ourselves, instead of allowing the data to choose the questions for us, we can then select the data needed to answer them. If science is the act of building and testing theory, we can apply this theory to build monitoring programs. For example, if a model of a system indicates that a particular parameter (e.g., early-stage survival) strongly influences system dynamics, then that parameter might be the key aspect to measure in the field.

However, simply asking individuals to design better studies overlooks the systemic barriers at play. Researchers feel pressure not only to publish more, but also to publish papers with broad appeal in high-impact journals—often emphasizing large-scale or global datasets. There is a false and pervasive narrative that more data necessarily means more insights. As others have called for, this means shifting academic hiring and promotion practices away from a focus on the number of publications or the flashiness of the journals. Counting papers and citations is simple; valuing them correctly is not.

Short grant cycles further constrain the kinds of ecological questions that scientists can pursue. Except for long-term initiatives such as the US National Science Foundation's National Ecological Observatory Network (NEON) and the Long Term Ecological Research (LTER) program, it remains difficult to sustain consistent monitoring efforts. The result is a patchwork of short-term, small-scale studies that may be underpowered or poorly coordinated across sites. Developing funding mechanisms that support long-term continuity, and embedding monitoring within institutional frameworks, would allow us to design programs that are more intentional, comparable, and resilient over time.

Satellites, sensors, and scientists will continue to collect data. But at times we must pause to ask whether we're still monitoring what matters—or simply what's convenient. Unless we reshape our questions, the data will continue to shape them for us. Building systems that reward thoughtful, long-term, and theory-informed monitoring will ensure that the era of big data becomes an era of better science—and better decisions.