



Redefining Ecological Science Using Data

JAMES R. HUNT

University of California, Berkeley, and the Berkeley Water Center

DENNIS D. BALDOCCHI

University of California, Berkeley

CATHARINE VAN INGEN

Microsoft Research

ECOLOGY IS THE STUDY OF LIFE and its interactions with the physical environment. Because climate change requires rapid adaptation, new data analysis tools are essential to quantify those changes in the midst of high natural variability. Ecology is a science in which studies have been performed primarily by small groups of individuals, with data recorded and stored in notebooks. But large synthesis studies are now being attempted by collaborations involving hundreds of scientists. These larger efforts are essential because of two developments: one in how science is done and the other in the resource management questions being asked. While collaboration synthesis studies are still nascent, their ever-increasing importance is clear. Computational support is integral to these collaborations and key to the scientific process.

HOW GLOBAL CHANGES ARE CHANGING ECOLOGICAL SCIENCE

The global climate and the Earth's landscape are changing, and scientists must quantify significant linkages between atmospheric, oceanic, and terrestrial processes to properly study the phenomena. For example, scientists are now asking how climate fluctuations in temperature, precipitation, solar radiation, length of growing season, and extreme weather events such as droughts affect the net carbon exchange between vegetation and the atmo-

sphere. This question spans many Earth science disciplines with their respective data, models, and assumptions.

These changes require a new approach to resolving resource management questions. In the short run of the next few decades, ecosystems cannot be restored to their former status. For example, with a warming climate on the West Coast of the United States, can historical data from coastal watersheds in southern California be used to predict the fish habitats of northern California coastal watersheds? Similarly, what can remote sensing tell us about deforestation? Addressing these challenges requires a synthesis of data and models that spans length scales from the very local (river pools) to the global (oceanic circulations) and spans time scales from a few tens of milliseconds to centuries.

AN EXAMPLE OF ECOLOGICAL SYNTHESIS

Figure 1 shows a simple “science mash-up” example of a synthesis study. The graph compares annual runoff from relatively small watersheds in the foothills of the Sierra Nevada in California to local annual precipitation over multiple years. Annual runoff values were obtained from the U.S. Geological Survey (USGS) for three of the gauging stations along Dry Creek and the Schubert University of California experimental field site.¹ Long-term precipitation records from nearby rain gauges were obtained from the National Climatic Data Center.² The precipitation that does not run off undergoes evapotranspiration (ET) that is largely dominated by watershed vegetation. In these watersheds, a single value of 400 mm is observed over all years of data. A similar value of annual ET was obtained by independent

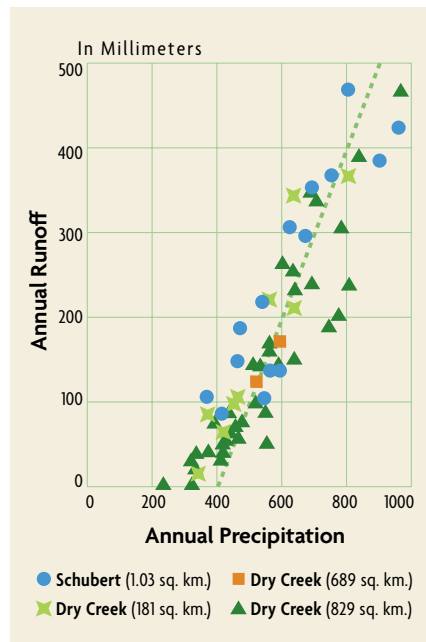


FIGURE 1. Simple annual water balance to estimate evapotranspiration in Sierra Nevada foothill watersheds. The dashed line represents an annual ET of 400 mm.

¹ <http://waterdata.usgs.gov/nwis>

² www.ncdc.noaa.gov

measurement from atmospheric sensors deployed over an oak savannah ecosystem at the AmeriFlux Tonzi Ranch tower.³ This synthesis of historical data defines a watershed model appropriate for historical conditions and provides a reference frame for addressing climate change effects in a highly variable system.

THE COMING FLOOD OF ECOLOGICAL DATA

These new synthesis studies are enabled by the confluence of low-cost sensors, remote sensing, Internet connectivity, and commodity computing. Sensor deployments by research groups are shifting from short campaigns to long-term monitoring with finer-scale and more diverse instruments. Satellites give global coverage particularly to remote or harsh regions where field research is hampered by physical and political logistics. Internet connectivity is enabling data sharing across organizations and disciplines. The result of these first three factors is a data flood. Commodity computing provides part of the solution, by allowing for the flood to be paired with models that incorporate different physical and biological processes and allowing for different models to be linked to span the length and time scales of interest.

The flood of ecological data and ecological science synthesis presents unique computing infrastructure challenges and new opportunities. Unlike sciences such as physics or astronomy, in which detectors are shared, in ecological science data are generated by a wide variety of groups using a wide variety of sampling or simulation methodologies and data standards. As shown earlier in Figure 1, the use of published data from two different sources was essential to obtain evapotranspiration. This synthesis required digital access to long records, separate processing of those datasets to arrive at ET, and finally verification with independent flux tower measurements. Other synthetic activities will require access to evolving resources from government organizations such as NASA or USGS, science collaborations such as the National Ecological Observatory Network and the WATERS Network,⁴ individual university science research groups such as Life Under Your Feet,⁵ and even citizen scientist groups such as the Community Collaborative Rain, Hail and Snow Network⁶ and the USA National Phenology Network.⁷

While the bulk of the data start out as digital, originating from the field sensor,

³ www.fluxdata.org:8080/SitePages/siteInfo.aspx?US-Ton

⁴ www.watersnet.org

⁵ www.lifeunderyourfeet.org

⁶ www.cocorahs.org

⁷ www.usanpn.org

radar, or satellite, the historic data and field data, which are critical for the science, are being digitized. The latter data are not always evenly spaced time series; they can include the date of leaf budding, or aerial imagery at different wavelengths and resolutions to assess quantities throughout the watershed such as soil moisture, vegetation, and land use. Deriving science variables from remote sensing remains an active area of research; as such, hard-won field measurements often form the ground truth necessary to develop conversion algorithms. Citizen science field observations such as plant species, plant growth (budding dates or tree ring growth, for example), and fish and bird counts are becoming increasingly important. Integrating such diverse information is an ever-increasing challenge to science analysis.

NAVIGATING THE ECOLOGICAL DATA FLOOD

The first step in any ecological science analysis is data discovery and harmonization. Larger datasets are discoverable today; smaller and historic datasets are often found by word of mouth. Because of the diversity of data publishers, no single reporting protocol exists. Unit conversions, geospatial reprojections, and time/length scale regularizations are a way of life. Science data catalog portals such as SciScope⁸ and Web services with common data models such as those from the Open Geospatial Consortium⁹ are evolving.

Integral to these science data search portals is knowledge of geospatial features and variable namespace mediation. The first enables searches across study watersheds or geological regions as well as simple polygon bounding boxes. The second enables searches to include multiple search terms—such as “rainfall,” “precipitation,” and “precip”—when searching across data repositories with different naming conventions. A new generation of metadata registries that use semantic Web technologies will enable richer searches as well as automated name and unit conversions. The combination of both developments will enable science data searches such as “Find me the daily river flow and suspended sediment discharge data from all watersheds in Washington State with more than 30 inches of annual rainfall.”

MOVING ECOLOGICAL SYNTHESIS INTO THE CLOUD

Large synthesis datasets are also leading to a migration from the desktop to cloud computing. Most ecological science datasets have been collections of files. An example is the Fluxnet LaThuile synthesis dataset, containing 966 site-years of sensor

⁸ www.sciscope.org

⁹ www.opengeospatial.org

data from 253 sites around the world. The data for each site-year is published as a simple comma-separated or MATLAB-ready file of either daily aggregates or half-hourly aggregates. Most of the scientists download some or all of the files and then perform analyses locally. Other scientists are using an alternative cloud service that links MATLAB on the desktop to a SQL Server Analysis Services data cube in the cloud. The data appears local, but the scientists need not be bothered with the individual file handling. Local download and manipulation of the remote sensing data that would complement that sensor data are not practical for many scientists. A cloud analysis now in progress using both to compute changes in evapotranspiration across the United States over the last 10 years will download 3 terabytes of imagery and use 4,000 CPU hours of processing to generate less than 100 MB of results. Doing the analysis off the desktop leverages the higher bandwidth, large temporary storage capacity, and compute farm available in the cloud.

Synthesis studies also create a need for collaborative tools in the cloud. Science data has value for data-owner scientists in the form of publications, grants, reputation, and students. Sharing data with others should increase rather than decrease that value. Determining the appropriate citations, acknowledgment, and/or co-authorship policies for synthesis papers remains an open area of discussion in larger collaborations such as Fluxnet¹⁰ and the North American Carbon Program.¹¹ Journal space and authorship limitations are an important concern in these discussions. Addressing the ethical question of what it means to be a co-author is essential: Is contributing data sufficient when that contribution is based on significant intellectual and physical effort? Once such policies are agreed upon, simple collaborative tools in the cloud can greatly reduce the logistics required to publish a paper, provide a location for the discovery of collaboration authors, and enable researchers to track how their data are used.

HOW CYBERINFRASTRUCTURE IS CHANGING ECOLOGICAL SCIENCE

The flood of ecological data will break down scientific silos and enable a new generation of scientific research. The goal of understanding the impacts of climate change is driving research that spans disciplines such as plant physiology, soil science, meteorology, oceanography, hydrology, and fluvial geomorphology. Bridging the diverse length and time scales involved will require a collection of cooperating models. Synthesizing the field observations with those model results at key length

¹⁰ www.fluxdata.org

¹¹ www.nacarbon.org/nacp

and time scales is crucial to the development and validation of such models.

The diversity of ecological dataset size, dataset semantics, and dataset publisher concerns poses a cyberinfrastructure challenge that will be addressed over the next several years. Synthesis science drives not only direct conversations but also virtual ones between scientists of different backgrounds. Advances in metadata representation can break down the semantic and syntactic barriers to those conversations. Data visualizations that range from our simple mashup to more complex virtual worlds are also key elements in those conversations. Cloud access to discoverable, distributed datasets and, perhaps even more important, enabling cloud data analyses near the more massive datasets will enable a new generation of cross-discipline science.