



1. EARTH AND ENVIRONMENT





Introduction

DAN FAY | Microsoft Research

CHANGE IS INEVITABLE—the Universe expands, nature adapts and evolves, and so must the scientific tools and technologies that we employ to feed our unrelenting quest for greater knowledge in space, Earth, and environmental sciences. The opportunities and challenges are many. New computing technologies such as cloud computing and multicore processors cannot provide the entire solution in their generic forms. But effective and timely application of such technologies can help us significantly advance our understanding of our world, including its environmental challenges and how we might address them.

With science moving toward being computational and data based, key technology challenges include the need to better capture, analyze, model, and visualize scientific information. The ultimate goal is to aid scientists, researchers, policymakers, and the general public in making informed decisions. As society demands action and responsiveness to growing environmental issues, new types of applications grounded in scientific research will need to move from raw discovery and eliciting basic data that leads to knowledge to informing practical decisions. Active issues such as climate change will not wait until scientists have all the data to fill their knowledge gaps.

As evidenced by the articles in this part of the book, scientists are indeed actively pursuing scientific understanding through the

use of new computing technologies. Szalay and Blakeley describe Jim Gray's informal rules for data-centric development and how they serve as a blueprint for making large-scale datasets available through the use of databases, leveraging the built-in data management as well as the parallel processing inherent in SQL servers.

In order to facilitate informed decisions based on reliable scientific evidence, Dozier and Gail explore how the applied use of technology and current scientific knowledge is key to providing tools to policy and decision makers. Hunt, Baldocchi, and van Ingen describe the changes under way in ecological science in moving from "science in the small" to large collaborations based on synthesis of data. These aggregated datasets expose the need for collaborative tools in the cloud as well as easy-to-use visualization and analysis tools. Delaney and Barga then provide compelling insights into the need for real-time monitoring of the complex dynamics in the sea by creating an interactive ocean laboratory. This novel cyberinfrastructure will enable new discoveries and insights through improved ocean models.

The need for novel scientific browsing technologies is highlighted by Goodman and Wong. To advance the linkage across existing resources, astronomers can use a new class of visualization tools, such as the WorldWide Telescope (WWT). This new class of tool offers access to data and information not only to professional scientists but also the general public, both for education and possibly to enable new discoveries by anyone with access to the Internet. Finally, Lehning et al. provide details about the use of densely deployed real-time sensors combined with visualization for increased understanding of environmental dynamics—like a virtual telescope looking back at the Earth. These applications illustrate how scientists and technologists have the opportunity to embrace and involve citizen scientists in their efforts.

In Part 1 and throughout the book, we see new sensors and infrastructures enabling real-time access to potentially enormous quantities of data, but with experimental repeatability through the use of workflows. Service-oriented architectures are helping to mitigate the transition to new underlying technologies and enable the linkage of data and resources. This rapidly evolving process is the only mechanism we have to deal with the data deluge arising from our instruments.

The question before us is how the world's intellectual and technological resources can be best orchestrated to authoritatively guide our responses to current and future societal challenges. The articles that follow provide some great answers.



Gray's Laws: Database-centric Computing in Science

**ALEXANDER S.
SZALAY**

The Johns Hopkins
University

JOSÉ A. BLAKELEY

Microsoft

THE EXPLOSION IN SCIENTIFIC DATA has created a major challenge for cutting-edge scientific projects. With datasets growing beyond a few tens of terabytes, scientists have no off-the-shelf solutions that they can readily use to manage and analyze the data [1]. Successful projects to date have deployed various combinations of flat files and databases [2]. However, most of these solutions have been tailored to specific projects and would not be easy to generalize or scale to the next generation of experiments. Also, today's computer architectures are increasingly imbalanced; the latency gap between multi-core CPUs and mechanical hard disks is growing every year, making the challenges of data-intensive computing harder to overcome [3]. What is needed is a systematic and general approach to these problems with an architecture that can scale into the future.

GRAY'S LAWS

Jim Gray formulated several informal rules—or laws—that codify how to approach data engineering challenges related to large-scale scientific datasets. The laws are as follows:

1. Scientific computing is becoming increasingly data intensive.
2. The solution is in a “scale-out” architecture.
3. Bring computations to the data, rather than data to the computations.

4. Start the design with the “20 queries.”
5. Go from “working to working.”

It is important to realize that the analysis of observational datasets is severely limited by the relatively low I/O performance of most of today’s computing platforms. High-performance numerical simulations are also increasingly feeling the “I/O bottleneck.” Once datasets exceed the random access memory (RAM) capacity of the system, locality in a multi-tiered cache no longer helps [4]. Yet very few high-end platforms provide a fast enough I/O subsystem.

High-performance, scalable numerical computation also presents an algorithmic challenge. Traditional numerical analysis packages have been designed to operate on datasets that fit in RAM. To tackle analyses that are orders of magnitude larger, these packages must be redesigned to work in a multi-phase, divide-and-conquer manner while maintaining their numerical accuracy. This suggests an approach in which a large-scale problem is decomposed into smaller pieces that can be solved in RAM, whereas the rest of the dataset resides on disk. This approach is analogous to the way in which database algorithms such as sorts or joins work on datasets larger than RAM. These challenges are reaching a critical stage.

Buying larger network storage systems and attaching them to clusters of compute nodes will not solve the problem because network/interconnect speeds are not growing fast enough to cope with the yearly doubling of the necessary storage. Scale-out solutions advocate simple building blocks in which the data is partitioned among nodes with locally attached storage [5]. The smaller and simpler these blocks are, the better the balance between CPUs, disks, and networking can become. Gray envisaged simple “CyberBricks” where each disk drive has its own CPU and networking [6]. While the number of nodes on such a system would be much larger than in a traditional “scale-up” architecture, the simplicity and lower cost of each node and the aggregate performance would more than make up for the added complexity. With the emergence of solid-state disks and low-power motherboards, we are on the verge of being able to build such systems [7].

DATABASE-CENTRIC COMPUTING

Most scientific data analyses are performed in hierarchical steps. During the first pass, a subset of the data is extracted by either filtering on certain attributes (e.g., removing erroneous data) or extracting a vertical subset of the columns. In the next step, data are usually transformed or aggregated in some way. Of course, in more

complex datasets, these patterns are often accompanied by complex joins among multiple datasets, such as external calibrations or extracting and analyzing different parts of a gene sequence [8]. As datasets grow ever larger, the most efficient way to perform most of these computations is clearly to move the analysis functions as close to the data as possible. It also turns out that most of these patterns are easily expressed by a set-oriented, declarative language whose execution can benefit enormously from cost-based query optimization, automatic parallelism, and indexes.

Gray and his collaborators have shown on several projects that existing relational database technologies can be successfully applied in this context [9]. There are also seamless ways to integrate complex class libraries written in procedural languages as an extension of the underlying database engine [10, 11].

MapReduce has become a popular distributed data analysis and computing paradigm in recent years [12]. The principles behind this paradigm resemble the distributed grouping and aggregation capabilities that have existed in parallel relational database systems for some time. New-generation parallel database systems such as Teradata, Aster Data, and Vertica have rebranded these capabilities as “MapReduce in the database.” New benchmarks comparing the merits of each approach have been developed [13].

CONNECTING TO THE SCIENTISTS

One of the most challenging problems in designing scientific databases is to establish effective communication between the builder of the database and the domain scientists interested in the analysis. Most projects make the mistake of trying to be “everything for everyone.” It is clear that that some features are more important than others and that various design trade-offs are necessary, resulting in performance trade-offs.

Jim Gray came up with the heuristic rule of “20 queries.” On each project he was involved with, he asked for the 20 most important questions the researchers wanted the data system to answer. He said that five questions are not enough to see a broader pattern, and a hundred questions would result in a shortage of focus. Since most selections involving human choices follow a “long tail,” or so-called 1/f distribution, it is clear that the relative information in the queries ranked by importance is logarithmic, so the gain realized by going from approximately 20 ($2^{4.5}$) to 100 ($2^{6.5}$) is quite modest [14].

The “20 queries” rule is a moniker for a design step that engages the domain scientist and the database engineer in a conversation that helps bridge the semantic

gap between nouns and verbs used in the scientific domain and the entities and relationships stored in the database. Queries define the precise set of questions in terms of entities and relationships that domain scientists expect to pose to the database. At the end of a full iteration of this exercise, the domain scientist and the database speak a common language.

This approach has been very successful in keeping the design process focused on the most important features the system must support, while at the same time helping the domain scientists understand the database system trade-offs, thereby limiting “feature creep.”

Another design law is to move from working version to working version. Gray was very much aware of how quickly data-driven computing architecture changes, especially if it involves distributed data. New distributed computing paradigms come and go every other year, making it extremely difficult to engage in a multi-year top-down design and implementation cycle. By the time such a project is completed, the starting premises have become obsolete. If we build a system that starts working only if every one of its components functions correctly, we will never finish.

The only way to survive and make progress in such a world is to build modular systems in which individual components can be replaced as the underlying technologies evolve. Today’s service-oriented architectures are good examples of this. Web services have already gone through several major evolutionary stages, and the end is nowhere in sight.

FROM TERASCALE TO PETASCALE SCIENTIFIC DATABASES

By using Microsoft SQL Server, we have successfully tackled several projects on a scale from a few terabytes (TB) to tens of terabytes [15-17]. Implementing databases that will soon exceed 100 TB also looks rather straightforward [18], but it is not entirely clear how science will cross the petascale barrier. As databases become larger and larger, they will inevitably start using an increasingly scaled-out architecture. Data will be heavily partitioned, making distributed, non-local queries and distributed joins increasingly difficult.

For most of the petascale problems today, a simple data-crawling strategy over massively scaled-out, share-nothing data partitions has been adequate (MapReduce, Hadoop, etc.). But it is also clear that this layout is very suboptimal when a good index might provide better performance by orders of magnitude. Joins between tables of very different cardinalities have been notoriously difficult to use with these crawlers.

Databases have many things to offer in terms of more efficient plans. We also need to rethink the utility of expecting a monolithic result set. One can imagine crawlers over heavily partitioned databases implementing a construct that can provide results one bucket at a time, resulting in easier checkpointing and recovery in the middle of an extensive query. This approach is also useful for aggregate functions with a clause that would stop when the result is estimated to be within, for example, 99% accuracy. These simple enhancements would go a long way toward sidestepping huge monolithic queries—breaking them up into smaller, more manageable ones.

Cloud computing is another recently emerging paradigm. It offers obvious advantages, such as co-locating data with computations and an economy of scale in hosting the services. While these platforms obviously perform very well for their current intended use in search engines or elastic hosting of commercial Web sites, their role in scientific computing is yet to be clarified. In some scientific analysis scenarios, the data needs to be close to the experiment. In other cases, the nodes need to be tightly integrated with a very low latency. In yet other cases, very high I/O bandwidth is required. Each of these analysis strategies would be suboptimal in current virtualization environments. Certainly, more specialized data clouds are bound to emerge soon. In the next few years, we will see if scientific computing moves from universities to commercial service providers or whether it is necessary for the largest scientific data stores to be aggregated into one.

CONCLUSIONS

Experimental science is generating vast volumes of data. The Pan-STARRS project will capture 2.5 petabytes (PB) of data each year when in production [18]. The Large Hadron Collider will generate 50 to 100 PB of data each year, with about 20 PB of that data stored and processed on a worldwide federation of national grids linking 100,000 CPUs [19]. Yet generic data-centric solutions to cope with this volume of data and corresponding analyses are not readily available [20].

Scientists and scientific institutions need a template and collection of best practices that lead to balanced hardware architectures and corresponding software to deal with these volumes of data. This would reduce the need to reinvent the wheel. Database features such as declarative, set-oriented languages and automatic parallelism, which have been successful in building large-scale scientific applications, are clearly needed.

We believe that the current wave of databases can manage at least another order of magnitude in scale. So for the time being, we can continue to work. However,

it is time to start thinking about the next wave. Scientific databases are an early predictor of requirements that will be needed by conventional corporate applications; therefore, investments in these applications will lead to technologies that will be broadly applicable in a few years. Today's science challenges are good representatives of the data management challenges for the 21st century. Gray's Laws represent an excellent set of guiding principles for designing the data-intensive systems of the future.

REFERENCES

- [1] A. S. Szalay and J. Gray, "Science in an Exponential World," *Nature*, vol. 440, pp. 23–24, 2006, doi: 10.1038/440413a.
- [2] J. Becla and D. Wang, "Lessons Learned from Managing a Petabyte," CIDR 2005 Conference, Asilomar, 2005, doi: 10.2172/839755.
- [3] G. Bell, J. Gray, and A. Szalay, "Petascale Computational Systems: Balanced Cyber-Infrastructure in a Data-Centric World," *IEEE Computer*, vol. 39, pp. 110–112, 2006, doi: 10.1109/MC.2006.29.
- [4] W. W. Hsu and A. J. Smith, "Characteristics of I/O traffic in personal computer and server workloads," *IBM Sys. J.*, vol. 42, pp. 347–358, 2003, doi: 10.1147/sj.422.0347.
- [5] A. Szalay, G. Bell, et al., "GrayWulf: Scalable Clustered Architecture for Data Intensive Computing," Proc. HICSS-42 Conference, Hawaii, 2009, doi: 10.1109/HICSS.2009.750.
- [6] J. Gray, Cyberbricks Talk at DEC/NT Wizards Conference, 2004; T. Barclay, W. Chong, and J. Gray, "TerraServer Bricks – A High Availability Cluster Alternative," Microsoft Technical Report, MSR-TR-2004-107, http://research.microsoft.com/en-us/um/people/gray/talks/DEC_Cyberbrick.ppt.
- [7] A. S. Szalay, G. Bell, A. Terzis, A. S. White, and J. Vandenberg, "Low Power Amdahl Blades for Data-Intensive Computing," <http://perspectives.mvdirona.com/content/binary/AmdahlBladesV3.pdf>.
- [8] U. Roehm and J. A. Blakeley, "Data Management for High-Throughput Genomics," *Proc. CIDR*, 2009.
- [9] J. Gray, D. T. Liu, M. A. Nieto-Santisteban, A. S. Szalay, G. Heber, and D. DeWitt, "Scientific Data Management in the Coming Decade," *ACM SIGMOD Record*, vol. 34, no. 4, pp. 35–41, 2005; also MSR-TR-2005-10, doi: 10.1145/1107499.1107503.
- [10] A. Acheson et al., "Hosting the .NET Runtime in Microsoft SQL Server," ACM SIGMOD Conf., 2004, doi: 10.1145/1007568.1007669.
- [11] J. A. Blakeley, M. Henaire, C. Kleinerman, I. Kunen, A. Prout, B. Richards, and V. Rao, ".NET Database Programmability and Extensibility in Microsoft SQL Server," ACM SIGMOD Conf., 2008, doi: 10.1145/1376616.1376725.
- [12] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," OSDI, 2004, doi: 10.1145/1327452.1327492.
- [13] A. Pavlo et al., "A Comparison of Approaches to Large-Scale Data Analysis," ACM SIGMOD Conf., 2009, doi: 10.1145/1559845.1559865.
- [14] C. Anderson. *The Long Tail*. New York: Random House, 2007.
- [15] A. R. Thakar, A. S. Szalay, P. Z. Kunszt, and J. Gray, "The Sloan Digital Sky Survey Science Archive: Migrating a Multi-Terabyte Astronomical Archive from Object to Relational DBMS," *Comp. Sci. and Eng.*, vol. 5, no. 5, pp. 16–29, Sept. 2003.

- [16] A. Terzis, R. Musaloiu-E., J. Cogan, K. Szlavecz, A. Szalay, J. Gray, S. Ozer, M. Liang, J. Gupchup, and R. Burns, "Wireless Sensor Networks for Soil Science," *Int. J. Sensor Networks*, to be published 2009.
- [17] Y. Li, E. Perlman, M. Wan, Y. Yang, C. Meneveau, R. Burns, S. Chen, A. Szalay, and G. Eyink, "A public turbulence database cluster and applications to study Lagrangian evolution of velocity increments in turbulence," *J. Turbul.*, vol. 9, no. 31, pp. 1–29, 2008, doi: 10.1080/14685240802376389.
- [18] Pan-STARRS: Panoramic Survey Telescope and Rapid Response System, <http://pan-starrs.ifa.hawaii.edu>.
- [19] A. M. Parker, "Understanding the Universe," in *Towards 2020 Science*, Microsoft Corporation, 2006, http://research.microsoft.com/towards2020science/background_overview.htm.
- [20] G. Bell, T. Hey, and A. Szalay, "Beyond the Data Deluge," *Science*, vol. 323, no. 5919, pp. 1297–1298, 2009, doi: 10.1126/science.1170411.



The Emerging Science of Environmental Applications

JEFF DOZIER

University of California,
Santa Barbara

WILLIAM B. GAIL

Microsoft

THE SCIENCE OF EARTH AND ENVIRONMENT has matured through two major phases and is entering a third. In the first phase, which ended two decades ago, Earth and environmental science was largely discipline oriented and focused on developing knowledge in geology, atmospheric chemistry, ecosystems, and other aspects of the Earth system. In the 1980s, the scientific community recognized the close coupling of these disciplines and began to study them as interacting elements of a single system. During this second phase, the paradigm of Earth system science emerged. With it came the ability to understand complex, system-oriented phenomena such as climate change, which links concepts from atmospheric sciences, biology, and human behavior. Essential to the study of Earth's interacting systems was the ability to acquire, manage, and make available data from satellite observations; in parallel, new models were developed to express our growing understanding of the complex processes in the dynamic Earth system [1].

In the emerging third phase, knowledge developed primarily for the purpose of scientific understanding is being complemented by knowledge created to target practical decisions and action. This new knowledge endeavor can be referred to as the *science of environmental applications*. Climate change provides the most prominent example of the importance of this shift. Until now, the

climate science community has focused on critical questions involving basic knowledge, from measuring the amount of change to determining the causes. With the basic understanding now well established, the demand for climate applications knowledge is emerging. How do we quantify and monitor total forest biomass so that carbon markets can characterize supply? What are the implications of regional shifts in water resources for demographic trends, agricultural output, and energy production? To what extent will seawalls and other adaptations to rising sea level impact coasts?

These questions are informed by basic science, but they raise additional issues that can be addressed only by a new science discipline focused specifically on applications—a discipline that integrates physical, biogeochemical, engineering, and human processes. Its principal questions reflect a fundamental curiosity about the nature of the world we live in, tempered by the awareness that a question’s importance scales with its relevance to a societal imperative. As Nobel laureate and U.S. Secretary of Energy Steven Chu has remarked, “We seek solutions. We don’t seek—dare I say this?—just scientific papers anymore” [2].

To illustrate the relationships between basic science and applications, consider the role of snowmelt runoff in water supplies. Worldwide, 1 billion people depend on snow or glacier melt for their water resources [3]. Design and operations of water systems have traditionally relied on historical measurements in a stationary climate, along with empirical relationships and models. As climates and land use change, populations grow and relocate, and our built systems age and decay, these empirical methods of managing our water become inaccurate—a conundrum characterized as “stationarity is dead” [4]. Snowmelt commonly provides water for competing uses: urban and agricultural supply, hydropower, recreation, and ecosystems. In many areas, both rainfall and snowfall occur, raising the concern that a future warmer climate will lead to a greater fraction of precipitation as rain, with the water arriving months before agricultural demand peaks and with more rapid runoff leading to more floods. In these mixed rain and snow systems, the societal need is: How do we sustain flood control and the benefits that water provides to humans and ecosystems when changes in the timing and magnitude of runoff are likely to render existing infrastructure inadequate?

The solution to the societal need requires a more fundamental, process-based understanding of the water cycle. Currently, historical data drive practices and decisions for flood control and water supply systems. Flood operations and reservoir flood capacity are predetermined by regulatory orders that are static, regardless

of the type of water year, current state of the snowpack, or risk of flood. In many years, early snowmelt is not stored because statistically based projections anticipate floods that better information might suggest cannot materialize because of the absence of snow. The more we experience warming, the more frequently this occurrence will impact the water supply [5]. The related science challenges are: (1) The statistical methods in use do not try to estimate the basin's water balance, and with the current measurement networks even in the U.S., we lack adequate knowledge of the amount of snow in the basins; (2) We are unable to partition the input between rain and snow, or to partition that rain or snow between evapotranspiration and runoff; (3) We lack the knowledge to manage the relationship between snow cover, forests, and carbon stocks; (4) Runoff forecasts that are not based on physical principles relating to snowmelt are often inaccurate; and (5) We do not know what incentives and institutional arrangements would lead to better management of the watershed for ecosystem services.

Generally, models do not consider these kinds of interactions; hence the need for a *science of environmental applications*. Its core characteristics differentiate it from the basic science of Earth and environment:

- **Need driven versus curiosity driven.** Basic science is question driven; in contrast, the new applications science is guided more by societal needs than scientific curiosity. Rather than seeking answers to questions, it focuses on creating the ability to seek courses of action and determine their consequences.
- **Externally constrained.** External circumstances often dictate when and how applications knowledge is needed. The creation of carbon trading markets will not wait until we fully quantify forest carbon content. It will happen on a schedule dictated by policy and economics. Construction and repair of the urban water infrastructure will not wait for an understanding of evolving rainfall patterns. Applications science must be prepared to inform actions subject to these external drivers, not according to academic schedules based on when and how the best knowledge can be obtained.
- **Consequential and recursive.** Actions arising from our knowledge of the Earth often change the Earth, creating the need for new knowledge about what we have changed. For example, the more we knew in the past about locations of fish populations, the more the populations were overfished; our original knowledge about them became rapidly outdated through our own actions. Applications sci-

ence seeks to understand not just those aspects of the Earth addressed by a particular use scenario, but also the consequences and externalities that result from that use scenario. A recent example is the shift of agricultural land to corn-for-ethanol production—an effort to reduce climate change that we now recognize as significantly stressing scarce water resources.

- **Useful even when incomplete.** As the snowpack example illustrates, actions are often needed despite incomplete data or partial knowledge. The difficulty of establishing confidence in the quality of our knowledge is particularly disconcerting given the loss of stationarity associated with climate change. New means of making effective use of partial knowledge must be developed, including robust inference engines and statistical interpretation.
- **Scalable.** Basic science knowledge does not always scale to support applications needs. The example of carbon trading presents an excellent illustration. Basic science tells us how to relate carbon content to measurements of vegetation type and density, but it does not give us the tools that scale this to a global inventory. New knowledge tools must be built to accurately create and update this inventory through cost-effective remote sensing or other means.
- **Robust.** The decision makers who apply applications knowledge typically have limited comprehension of how the knowledge was developed and in what situations it is applicable. To avoid misuse, the knowledge must be characterized in highly robust terms. It must be stable over time and insensitive to individual interpretations, changing context, and special conditions.
- **Data intensive.** Basic science is data intensive in its own right, but data sources that support basic science are often insufficient to support applications. Localized impacts with global extent, such as intrusion of invasive species, are often difficult for centralized projects with small numbers of researchers to ascertain. New applications-appropriate sources must be identified, and new ways of observing (including the use of communities as data gatherers) must be developed.

Each of these characteristics implies development of *new knowledge types* and *new tools for acquiring that knowledge*. The snowpack example illustrates what this requirement means for a specific application area. Four elements have recently come together that make deployment of a measurement and information system

that can support decisions at a scale of a large river basin feasible: (1) accurate, sustained satellite estimates of snow-covered area across an entire mountain range; (2) reliable, low-cost sensors and telemetry systems for snow and soil moisture; (3) social science data that complement natural and engineered systems data to enable analysis of human decision making; and (4) cyberinfrastructure advances to integrate data and deliver them in near real time.

For snow-dominated drainage basins, the highest-priority scientific challenge is to estimate the spatial distribution and heterogeneity of the *snow water equivalent*—i.e., the amount of water that would result if the snow were to melt. Because of wind redistribution of snow after it falls, snow on the ground is far more heterogeneous than rainfall, with several meters of differences within a 10 to 100 m distance. Heterogeneity in snow depth smoothes the daily runoff because of the variability of the duration of meltwater in the snowpack [6]; seasonally, it produces quasi-riparian zones of increased soil moisture well into the summer. The approach to estimating the snow water equivalent involves several tasks using improved data: (1) extensive validation of the satellite estimates of snow cover and its reflectivity, as Figure 1 on the next page shows; (2) using results from an energy balance reconstruction of snow cover to improve interpolation from more extensive ground measurements and satellite data [7]; (3) development of innovative ways to characterize heterogeneity [8]; and (4) testing the interpolated estimates with a spatially distributed runoff model [9]. The measurements would also help clarify the accuracy in precipitation estimates from regional climate models.

This third phase of Earth and environmental science will evolve over the next decade as the scientific community begins to pursue it. Weather science has already built substantial capability in applications science; the larger field of Earth science will need to learn from and extend those efforts. The need for basic science and further discovery will not diminish, but instead will be augmented and extended by this new phase. The questions to address are both practically important and intellectually captivating. Will our hydrologic forecasting skill decline as changes in precipitation diminish the value of statistics obtained from historic patterns? Where will the next big climate change issue arise, and what policy actions taken today could allow us to anticipate it?

Equally important is improving how we apply this knowledge in our daily lives. The Internet and mobile telephones, with their global reach, provide new ways to disseminate information rapidly and widely. Information was available to avoid much of the devastation from the Asian tsunami and Hurricane Katrina, but we

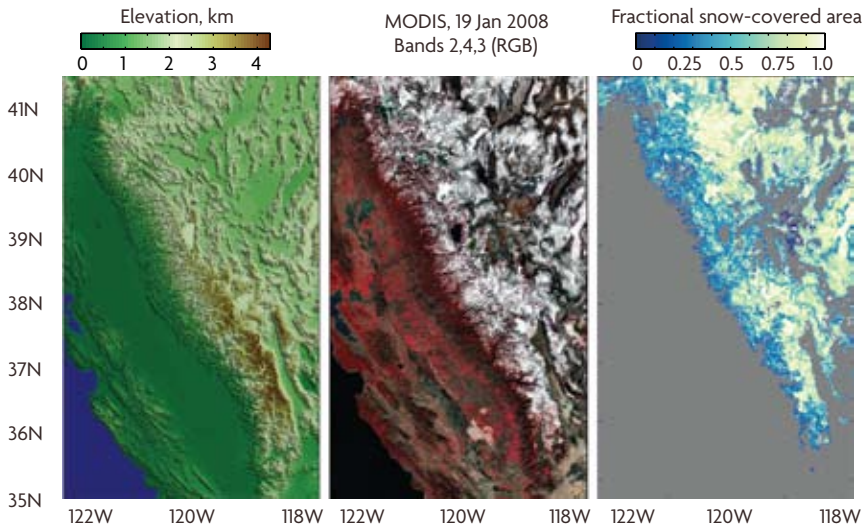


FIGURE 1. An illustration of the type of data that are useful in analyzing the snow cover. The left panel shows elevations of the Sierra Nevada and Central Valley of California, along with a portion of northwestern Nevada. The middle panel shows the raw satellite data in three spectral bands (0.841–0.876, 0.545–0.565, and 0.459–0.479 μm) from NASA’s Moderate Resolution Imaging Spectroradiometer (MODIS), which provides daily global data at 250 to 1000 m resolution in 36 spectral bands. From seven “land” bands at 500 m resolution, we derive the fractional snow-covered area—i.e., the fraction of each 500 m grid cell covered by snow, shown in the right panel [10].

lacked the tools for rapid decision making and communication of needed actions. Applications science is therefore integrative; it couples understanding of physical phenomena and research into the ways that people and organizations can use better knowledge to make decisions. The public as a whole can also become an important contributor to localized Earth observation, augmenting our limited satellite and sensor networks through devices as simple as mobile phone cameras. The ability to leverage this emerging data-gathering capability will be an important challenge for the new phase of environmental science.

The security and prosperity of nearly 7 billion people depend increasingly on our ability to gather and apply information about the world around us. Basic environ-

mental science has established an excellent starting point. We must now develop this into a robust science of environmental applications.

REFERENCES

- [1] National Research Council, *Earth Observations from Space: The First 50 Years of Scientific Achievement*. Washington, D.C.: National Academies Press, 2007.
- [2] R. DelVecchio, "UC Berkeley: Panel looks at control of emissions," *S.F. Chronicle*, March 22, 2007.
- [3] T. P. Barnett, J. C. Adam, and D. P. Lettenmaier, "Potential impacts of a warming climate on water availability in snow-dominated regions," *Nature*, vol. 438, pp. 303–309, 2005, doi: 10.1038/nature04141.
- [4] P. C. D. Milly, J. Betancourt, M. Falkenmark, R. M. Hirsch, Z. W. Kundzewicz, D. P. Lettenmaier, and R. J. Stouffer, "Stationarity is dead: whither water management?" *Science*, vol. 319, pp. 573–574, 2008, doi: 10.1126/science.1151915.
- [5] R. C. Bales, N. P. Molotch, T. H. Painter, M. D. Dettinger, R. Rice, and J. Dozier, "Mountain hydrology of the western United States," *Water Resour. Res.*, vol. 42, W08432, 2006, doi: 10.1029/2005WR004387.
- [6] J. D. Lundquist and M. D. Dettinger, "How snowpack heterogeneity affects diurnal streamflow timing," *Water Resour. Res.*, vol. 41, W05007, 2005, doi: 10.1029/2004WR003649.
- [7] D. W. Cline, R. C. Bales, and J. Dozier, "Estimating the spatial distribution of snow in mountain basins using remote sensing and energy balance modeling," *Water Resour. Res.*, vol. 34, pp. 1275–1285, 1998, doi: 10.1029/97WR03755.
- [8] N. P. Molotch and R. C. Bales, "Scaling snow observations from the point to the grid element: implications for observation network design," *Water Resour. Res.*, vol. 41, W11421, 2005, doi: 10.1029/2005WR004229.
- [9] C. L. Tague and L. E. Band, "RHESys: regional hydro-ecologic simulation system—an object-oriented approach to spatially distributed modeling of carbon, water, and nutrient cycling," *Earth Int.*, vol. 19, pp. 1–42, 2004.
- [10] T. H. Painter, K. Rittger, C. McKenzie, R. E. Davis, and J. Dozier, "Retrieval of subpixel snow-covered area, grain size, and albedo from MODIS," *Remote Sens. Environ.*, vol. 113, pp. 868–879, 2009, doi: 10.1016/j.rse.2009.01.001.



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

¹ <http://waterdata.usgs.gov/nwis>
² www.ncdc.noaa.gov

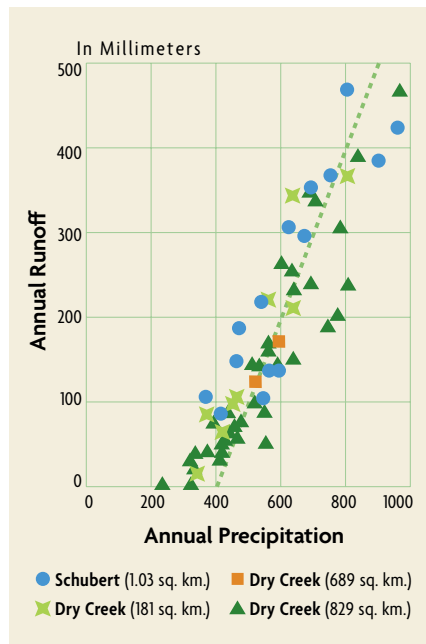


FIGURE 1. Simple annual water balance to estimate evapotranspiration in Sierra Nevada foothill watersheds. The dashed line represents an annual ET of 400 mm.

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.



A 2020 Vision for Ocean Science

JOHN R. DELANEY
University of Washington

ROGER S. BARGA
Microsoft Research

THE GLOBAL OCEAN is the last physical frontier on Earth. Covering 70 percent of the planetary surface, it is the largest, most complex biome we know. The ocean is a huge, mobile reservoir of heat and chemical mass. As such, it is the “engine” that drives weather-climate systems across the ocean basins and the continents, directly affecting food production, drought, and flooding on land. Water is effectively opaque to electromagnetic radiation, so the seafloor has not been as well mapped as the surfaces of Mars and Venus, and although the spatial relationships within the ocean basins are well understood to a first order, the long- and short-term temporal variations and the complexities of ocean dynamics are poorly understood.

The ultimate repository of human waste, the ocean has absorbed nearly half of the fossil carbon released since 1800. The ocean basins are a source of hazards: earthquakes, tsunamis, and giant storms. These events are episodic, powerful, often highly mobile, and frequently unpredictable. Because the ocean basins are a vast, but finite, repository of living and non-living resources, we turn to them for food, energy, and the many minerals necessary to sustain a broad range of human lifestyles. Many scientists believe that underwater volcanoes were the crucible in which early life began on Earth and perhaps on other planets. The oceans connect all continents; they are owned by no one, yet they belong

to all of us by virtue of their mobile nature. The oceans may be viewed as the common heritage of humankind, the responsibility and life support of us all.

OCEAN COMPLEXITY

Our challenge is to optimize the benefits and mitigate the risks of living on a planet dominated by two major energy sources: sunlight driving the atmosphere and much of the upper ocean, and internal heat driving plate tectonics and portions of the lower ocean. For more than 4 billion years, the global ocean has responded to and integrated the impacts of these two powerful driving forces as the Earth, the oceans, the atmosphere, and life have co-evolved. As a consequence, our oceans have had a long, complicated history, producing today's immensely complex system in which thousands of physical, chemical, and biological processes continually interact over many scales of time and space as the oceans maintain our planetary-scale ecological "comfort zone."

Figure 1 captures a small fraction of this complexity, which is constantly driven by energy from above and below. Deeper understanding of this "global life-support system" requires entirely novel research approaches that will allow broad spectrum, interactive ocean processes to be studied simultaneously and interactively by many scientists—approaches that enable continuous *in situ* examination of linkages among many processes in a coherent time and space framework. Implementing these powerful new approaches is both the challenge and the vision of next-generation ocean science.

HISTORICAL PERSPECTIVE

For thousands of years, humans have gone to sea in ships to escape, to conquer, to trade, and to explore. Between October 1957 and January 1960, we launched the first Earth-orbiting satellite and dove to the deepest part of the ocean. Ships, satellites, and submarines have been the mainstays of spatially focused oceanographic research and exploration for the past 50 years. We are now poised on the next threshold of technological breakthrough that will advance oceanic discovery; this time, exploration will be focused on the time domain and interacting processes. This new era will draw deeply on the emergence, and convergence, of many rapidly evolving new technologies. These changes are setting the scene for what Marcel Proust called "[t]he real voyage of discovery, [which] lies not in seeking new landscapes, but in having new eyes."

In many ways, this "vision" of next-generation oceanographic research and

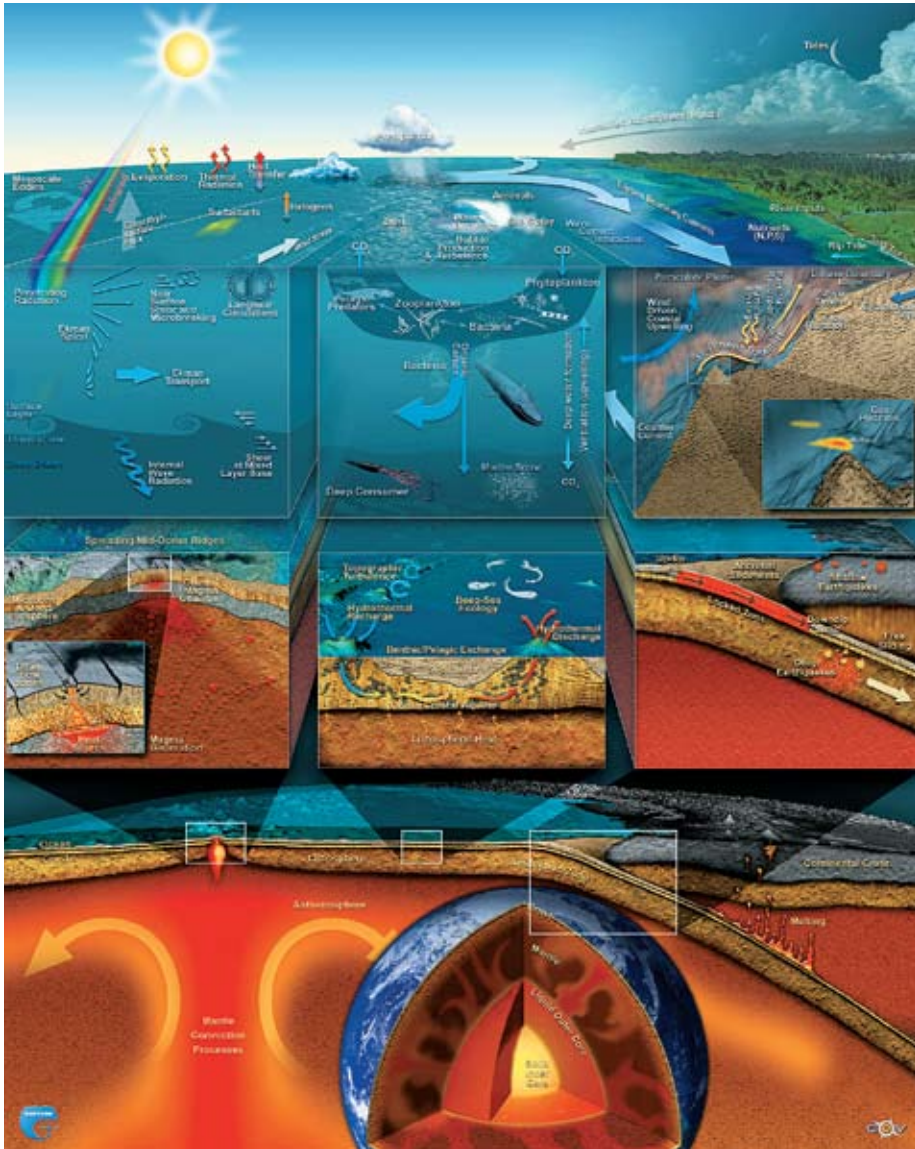


FIGURE 1.

Two primary energy sources powerfully influence the ocean basins: sunlight and its radiant energy, and internal heat with its convective and conductive input. Understanding the complexity of the oceans requires documenting and quantifying—in a well-defined time-space framework over decades—myriad processes that are constantly changing and interacting with one another.

Illustration designed by John Delaney and Mark Stoermer; created by the Center for Environmental Visualization (CEV) for the NEPTUNE Program.

education involves utilizing a wide range of innovative technologies to simultaneously and continuously “see,” or sense, many different processes operating throughout entire volumes of the ocean *from a perspective within the ocean*. Some of these same capabilities will enable remote *in situ* detection of critical changes taking place within selected ocean volumes. Rapid reconfiguration of key sensor arrays linked to the Internet via submarine electro-optical cables will allow us to capture, image, document, and measure energetic and previously inaccessible phenomena such as erupting volcanoes, major migration patterns, large submarine slumps, big earthquakes, giant storms, and a host of other complex phenomena that have been largely inaccessible to scientific study.

THE FOURTH PARADIGM

The ocean has been chronically under-sampled for as long as humans have been trying to characterize its innate complexity. In a very real sense, the current suite of computationally intensive numerical/theoretical models of ocean behavior has outstripped the requisite level of actual data necessary to ground those models in reality. As a consequence, we have been unable to even come close to useful predictive models of the real behavior of the oceans. Only by quantifying powerful episodic events, like giant storms and erupting volcanoes, within the context of longer-term decadal changes can we begin to approach dependable predictive models of ocean behavior. Over time, as the adaptive models are progressively refined by continual comparison with actual data flowing from real systems, we slowly gain the ability to predict the future behavior of these immensely complex natural systems. To achieve that goal, we must take steps to fundamentally change the way we approach oceanography.

This path has several crucial steps. We must be able to document conditions and measure fluxes *within the volume of the ocean, simultaneously and in real time*, over many scales of time and space, regardless of the depth, energy, mobility, or complexity of the processes involved. These measurements must be made using co-located arrays of many sensor types, operated by many investigators over periods of decades to centuries. And the data must be collected, archived, visualized, and compared immediately to model simulations that are explicitly configured to address complexity at scales comparable in time and space to the actual measurements.

This approach offers three major advantages: (1) The models must progressively emulate the measured reality through constant comparison with data to capture the real behavior of the oceans in “model space” to move toward more predictive

simulations; (2) When the models and the data disagree, assuming the data are valid, we must immediately adapt at-sea sensor-robot systems to fully characterize the events that are unfolding because they obviously offer new insights into the complexities we seek to capture in the failed models; (3) By making and archiving all observations and measurements in coherently indexed time and space frameworks, we can allow many investigators (even those not involved in the data collection) to examine correlations among any number of selected phenomena during, or long after, the time that the events or processes occur. If the archived data are immediately and widely available via the Internet, the potential for discovery rises substantially because of the growing number of potential investigators who can explore a rapidly expanding spectrum of “parameter space.” For scientists operating in this data-intensive environment, there will be a need for development of a new suite of scientific workflow products that can facilitate the archiving, assimilation, visualization, modeling, and interpretation of the information about all scientific systems of interest. Several workshop reports that offer examples of these “workflow products” are available in the open literature [1, 2].

EMERGENCE AND CONVERGENCE

Ocean science is becoming the beneficiary of a host of powerful *emergent* technologies driven by many communities that are entirely external to the world of ocean research—they include, but are not limited to, nanotechnology, biotechnology, information technology, computational modeling, imaging technologies, and robotics. More powerful yet will be the progressive *convergence* of these enabling capabilities as they are adapted to conduct sophisticated remote marine operations in novel ways by combining innovative technologies into appropriate investigative or experimental systems.

For example, computer-enabled support activities must include massive data storage systems, cloud computing, scientific workflow, advanced visualization displays, and handheld supercomputing. Instead of batteries and satellites being used to operate remote installations, electrical power and the vast bandwidth of optical fiber will be used to transform the kinds of scientific and educational activities that can be conducted within the ocean. Adaptation of industry-standard electro-optical cables for use in oceanographic research can fundamentally change the nature of human telepresence throughout the full volume of the oceans by introducing unprecedented but routinely available power and bandwidth into “ocean space.” High-resolution optical and acoustic sensing will be part of the broader technology

of “ocean imaging systems.” These approaches will include routine use of high-definition video, in stereo if needed, as well as high-resolution sonar, acoustic lenses, laser imaging, and volumetric sampling. Advanced sensor technologies will include chemical sensing using remote, and mobile, mass spectrometers and gas chromatographs, eco-genomic analysis, and adaptive sampling techniques.

AN INTEGRATED APPROACH

After decades of planning [3, 4], the U.S. National Science Foundation (NSF) is on the verge of investing more than US\$600 million over 6 years in the construction and early operation of an innovative infrastructure known as the Ocean Observatories Initiative (OOI) [4]. The design life of the program is 25 years. In addition to making much-needed high-latitude and coastal measurements supported by relatively low-bandwidth satellite communications systems, this initiative will include a transformative undertaking to implement electro-optically cabled observing systems in the northeast Pacific Ocean [5-7] off the coasts of Washington, Oregon, and British Columbia, as illustrated in Figure 2.¹

These interactive, distributed sensor networks in the U.S. and Canada will create a large-aperture “natural laboratory” for conducting a wide range of long-term innovative experiments within the ocean volume using real-time control over the entire “laboratory” system. Extending unprecedented power and bandwidth to a wide range of interactive sensors, instruments, and robots distributed throughout the ocean water, at the air-sea interface, on the seafloor, and below the seafloor within drill holes will empower next-generation creativity and exploration of the time domain among a broad spectrum of investigators. The University of Washington leads the cabled component of the NSF initiative, known as the Regional Scale Nodes (formerly known, and funded, as NEPTUNE); the University of Victoria leads the effort in Canada, known as NEPTUNE Canada. The two approaches were conceived jointly in 2000 as a collaborative U.S.-Canadian effort. The Consortium for Ocean Leadership in Washington, D.C., is managing and integrating the entire OOI system for NSF. Woods Hole Oceanographic Institution and the University of California, San Diego, are responsible for overseeing the Coastal-Global and Cyber-Infrastructure portions of the program, respectively. Oregon State University and Scripps Institution of Oceanography are participants in the Coastal-Global portion of the OOI.

¹ www.interactiveoceans.ocean.washington.edu

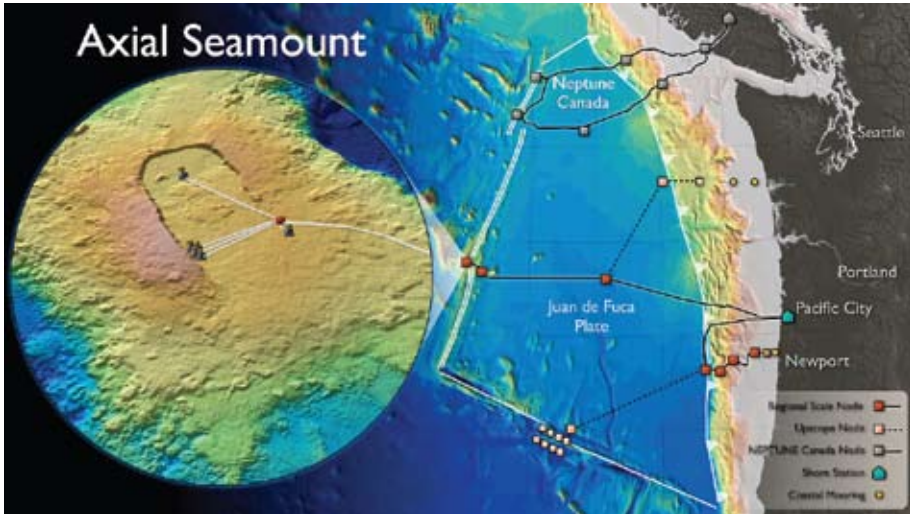


FIGURE 2.

A portion of the OOI focuses on the dynamic behavior of the Juan de Fuca Plate and the energetic processes operating in the overlying ocean and atmosphere. Recent modifications in the Regional Scale Nodes (RSN) have focused on delivery of the elements shown in red, and the pink components are future expansion. The inset shows the crest of Axial Seamount along the active Juan de Fuca Ridge. Each square block site will provide unprecedented electrical power and bandwidth available for research and education. Many of the processes shown in Figure 1 can be examined at the sites here.

Image created by CEV for OOI-RSN.

The cabled ocean observatory approach will revolutionize ocean science by providing interactive access to ocean data and instruments 24/7/365 over two to three decades. More than 1,200 kilometers of electro-optical submarine cable will deliver many tens of kilowatts of power to seafloor nodes, where instruments that might spread over a 50 km radius for each node will be plugged in directly or via secondary extension cables. The primary cable will provide between 2.5 and 10 gigabit/sec bandwidth connectivity between land and a growing number of fixed sensor packages and mobile sensor platforms. We expect that a host of novel approaches to oceanography will evolve based on the availability of *in situ* power and bandwidth. A major benefit will be the real-time data return and command-control of fleets of remotely operated vehicles (ROVs) and autonomous underwater vehicles



FIGURE 3.

Next-generation scientists or citizens. This virtual picture shows a deep ocean octopus, known as Grimpoteuthis, and a portion of a submarine hydrothermal system on the Juan de Fuca Ridge. Such real-time displays of 3-D HD video will be routine within 5 years.

Graphic designed by Mark Stoermer and created by CEV for NEPTUNE in 2005.

(AUVs). The infrastructure will be adaptable, expandable, and exportable to interested users. Data policy for the OOI calls for all information to be made available to all interested users via the Internet (with the exception of information bearing on national security).

Hardwired to the Internet, the cabled observatories will provide scientists, students, educators, and the public with virtual access to remarkable parts of our planet that are rarely visited by humans. In effect, the Internet will be extended to the seafloor, with the ability to interact with a host of instruments, including HD video live from the many environments within the oceans, as illustrated in Figure 3. The cabled observatory systems will be able to capture processes at the scale of the tectonic plate, mesoscale oceanic eddies, or even smaller scales. Research into representative activities responsible for climate change, major biological productivity at the base of the food chain, or encroaching ocean acidification (to name a few) will be readily conducted with this new infrastructure. Novel studies

of mid-ocean spreading centers, transform faults, and especially processes in the subduction zone at the base of the continental slope, which may trigger massive earthquakes in the Pacific Northwest, will also be addressable using the same investment in the same cabled infrastructure.

This interactive ocean laboratory will be enabled by a common cyberinfrastructure that integrates multiple observatories, thousands of instruments, tens of thousands of users, and petabytes of data. The goals of the cabled ocean observatory can be achieved only if the at-sea portion is complemented by state-of-the-art information technology infrastructure resulting from a strong collaborative effort between computer scientists and ocean scientists. Such collaboration will allow scientists to interact with the ocean through real-time command and control of sensors; provide models with a continuous data feed; automate data quality control and calibration; and support novel approaches to data management, analysis, and visualization.

WHAT IS POSSIBLE?

Figure 4 on the next page depicts some of the potentially transformative capabilities that could emerge in ocean science by 2020. In the long term, a key element of the introduction of unprecedented power and bandwidth for use within the ocean basins will be the potential for bold and integrative designs and developments that enhance our understanding of, and perhaps our ability to predict, the behavior of Earth, ocean, and atmosphere interactions and their bearing on a sustainable planetary habitat.

CONCLUSION

The cabled ocean observatory merges dramatic technological advancements in sensor technologies, robotic systems, high-speed communication, eco-genomics, and nanotechnology with ocean observatory infrastructure in ways that will substantially transform the approaches that scientists, educators, technologists, and policymakers take in interacting with the dynamic global ocean. Over the coming decades, most nations will implement systems of this type in the offshore extensions of their territorial seas. As these systems become more sophisticated and data become routinely available via the Internet, the Internet will emerge as the most powerful oceanographic research tool on the planet. In this fashion, the legacy of Jim Gray will continue to grow as we learn to discover truths and insights within the data we already have “in the can.”

While the cabled observatory will have profound ramifications for the manner

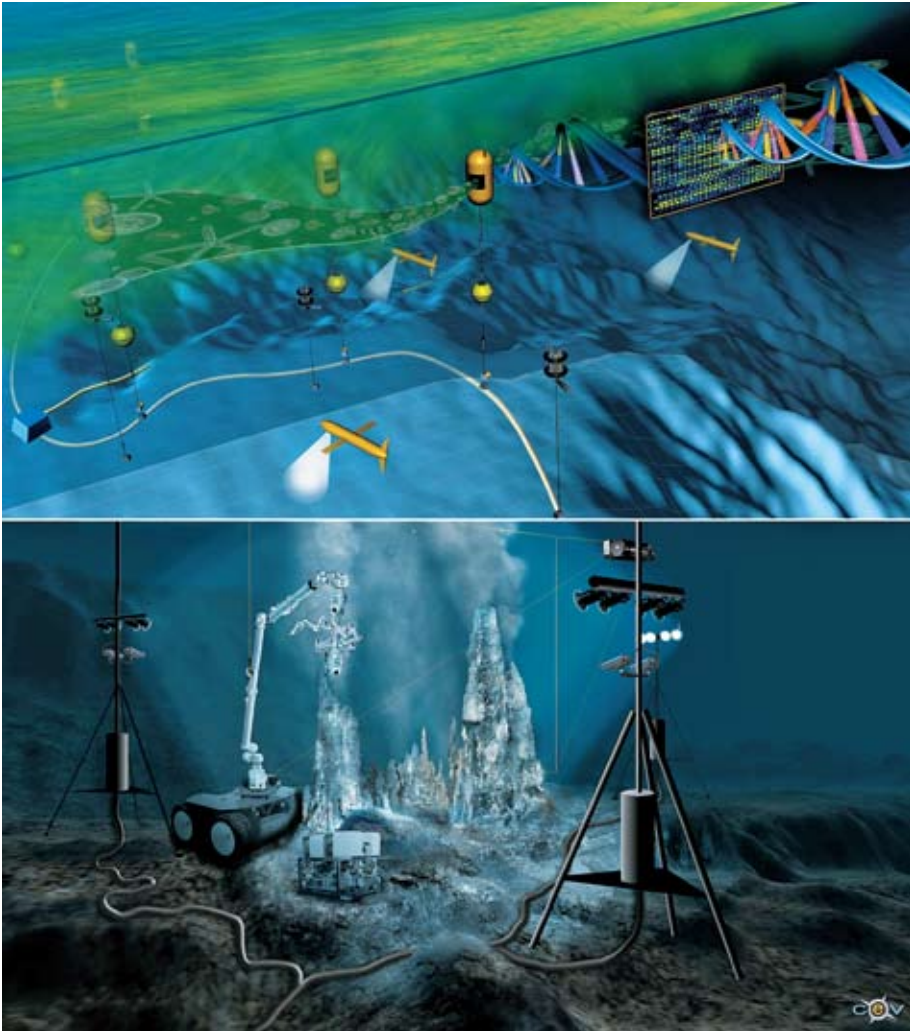


FIGURE 4.

Some of the transformative developments that could become routine within 5 years with the added power of a cabled support system. The top image shows miniaturized genomic analysis systems adapted from land laboratories to the ocean to allow scientists, with the flip of a switch in their lab hundreds of miles away, to sample ambient flow remotely and run in situ gene sequencing operations within the ocean. The data can be made available on the Internet within minutes of the decision to sample microbes in an erupting submarine volcanic plume or a seasonally driven phytoplankton bloom. The lower part shows a conceptual illustration of an entire remote analytical-biological laboratory on the seafloor that allows a variety of key measurements or dissections to be made in situ using stereo high-definition video to guide high-precision remote manipulations.

Scientific concepts by Ginger Armbrust and John Delaney; graphic design by Mark Stoermer for CEV.

in which scientists, engineers, and educators conduct their professional activities, the most far-reaching effects may be a significant shift in public attitudes toward the oceans as well as toward the scientific process. The real-time data and high-speed communications inherent in cabled remote observing systems will also open entirely new avenues for the public to interact with the natural world.

In the final analysis, having predictive models of how the ocean functions based on decades of refining sophisticated computer simulations against high-quality observations from distributed sensor networks will form the basis for learning to manage, or at least adapt to, the most powerful climate modulating system on the planet—the global ocean.

ACKNOWLEDGMENTS

We gratefully acknowledge the significant influence of Jim Gray, who unflinchingly stated that this cabled ocean observing approach using high-bandwidth and real-time data flow would be integral to human progress and understanding of the world we live in. We are also pleased to acknowledge the support of the University of Washington, the National Science Foundation, the Consortium for Ocean Leadership, and the Microsoft External Research group for technical collaboration and financial support. NSF and the National Oceanographic Partnership Program were particularly supportive of the early development of the NEPTUNE concept from 1998 to 2005, through grants to J. R. Delaney. Deborah Kelley, Nancy Penrose, and Mark Stoermer contributed significantly to the preparation of this manuscript and to conversations bearing on the content.

REFERENCES

- [1] “Project Trident: A Scientific Workflow Workbench Brings Clarity to Data,” <http://research.microsoft.com/en-us/collaboration/focus/e3/workflowtool.aspx>.
- [2] Two URLs for the NSF Workshop on Challenges of Scientific Workflows: <http://grids.ucs.indiana.edu/ptliupages/publications/IEEEComputer-gil.pdf> http://vtcpc.isi.edu/wiki/index.php/Main_Page.
- [3] National Research Council of the National Academies, *Enabling Ocean Research in the 21st Century: Implementation of a Network of Ocean Observatories*. Washington, D.C.: National Academies Press, 2003, p. 220.
- [4] “Ocean Observatories Initiative (OOI) Scientific Objectives and Network Design: A Closer Look,” 2007, <http://ooi.ocean.washington.edu/cruise/cruiseFile/show/40>. Ocean Leadership Web site for the Ocean Observatories Initiative: www.oceanleadership.org/programs-and-partnerships/ocean-observing/ooi.
- [5] J. R. Delaney, F. N. Spiess, S. C. Solomon, R. Hessler, J. L. Karsten, J. A. Baross, R. T. Holcomb, D. Norton, R. E. McDuff, F. L. Sayles, J. Whitehead, D. Abbott, and L. Olson, “Scientific rationale for establishing long-term ocean bottom observatory/laboratory systems,” in *Marine Minerals*:

Resource Assessment Strategies, P. G. Teleki, M. R. Dobson, J. R. Moor, and U. von Stackelberg, Eds., 1987, pp. 389–411.

- [6] J. R. Delaney, G. R. Heath, A. D. Chave, B. M. Howe, and H. Kirkham, “NEPTUNE: Real-time ocean and earth sciences at the scale of a tectonic plate,” *Oceanography*, vol. 13, pp. 71–83, 2000, doi: 10.1109/OCEANS.2001.968033.
- [7] A. D. Chave, B. St. Arnaud, M. Abbott, J. R. Delaney, R. Johnson, E. Lazowska, A. R. Maffei, J. A. Orcutt, and L. Smarr, “A management concept for ocean observatories based on web services,” *Proc. Oceans’04/Techno-Ocean’04*, Kobe, Japan, Nov. 2004, p. 7, doi: 10.1109/OCEANS.2004.1406486.



Bringing the Night Sky Closer: Discoveries in the Data Deluge

ALYSSA A. GOODMAN
Harvard University
CURTIS G. WONG
Microsoft Research

THROUGHOUT HISTORY, ASTRONOMERS have been accustomed to data falling from the sky. But our relatively newfound ability to store the sky’s data in “clouds” offers us fascinating new ways to access, distribute, use, and analyze data, both in research and in education. Here we consider three inter-related questions: (1) What trends have we seen, and will soon see, in the growth of image and data collection from telescopes? (2) How might we address the growing challenge of finding the proverbial needle in the haystack of this data to facilitate scientific discovery? (3) What visualization and analytic opportunities does the future hold?

TRENDS IN DATA GROWTH

Astronomy has a history of data collection stretching back at least to Stonehenge more than three millennia ago. Over time, the format of the information recorded by astronomers has changed, from carvings in stone to written records and hand-drawn illustrations to photographs to digital media.

While the telescope (c. 1600) and the opening up of the electromagnetic spectrum beyond wavelengths visible to the human eye (c. 1940) led to qualitative changes in the nature of astronomical investigations, they did not increase the volume of collected data nearly as much as did the advent of the Digital Age.

Charge-coupled devices (CCDs), which came into widespread use by the 1980s, and equivalent detectors at non-optical wavelengths became much more efficient than traditional analog media (such as photographic plates). The resulting rise in the rate of photon collection caused the ongoing (and potentially perpetually accelerating) increase in data available to astronomers. The increasing capabilities and plummeting price of the digital devices used in signal processing, data analysis, and data storage, combined with the expansion of the World Wide Web, transformed astronomy from an observational science into a digital and computational science.

For example, the Large Synoptic Survey Telescope (LSST), coming within the decade, will produce more data in its first year of operation—1.28 petabytes—than any other telescope in history by a significant margin. The LSST will accomplish this feat by using very sensitive CCDs with huge numbers of pixels on a relatively large telescope with very fast optics ($f/1.234$) and a wide field of view (9.6 square degrees), and by taking a series of many shorter exposures (rather than the traditional longer exposures) that can be used to study the temporal behavior of astronomical sources. And while the LSST, Pan-STARRS, and other coming astronomical mega-projects—many at non-optical wavelengths—will produce huge datasets covering the whole sky, other groups and individuals will continue to add their own smaller, potentially more targeted, datasets.

For the remainder of this article, we will assume that the challenge of managing this explosive growth in data will be solved (likely through the clever use of “cloud” storage and novel data structures), and we will focus instead on how to offer better tools and novel technical and social analytics that will let us learn more about our universe.

A number of emerging trends can help us find the “needles in haystacks” of data available over the Internet, including crowdsourcing, democratization of access via new browsing technologies, and growing computational power.

CROWDSOURCING

The Sloan Digital Sky Survey was undertaken to image, and measure spectra for, millions of galaxies. Most of the galaxy images had never been viewed by a human because they were automatically extracted from wide-field images reduced in an automated pipeline. To test a claim that more galaxies rotate in an anticlockwise direction than clockwise, the Sloan team used custom code to create a Web page that served up pictures of galaxies to members of the public willing to play the online Galaxy Zoo game, which consists primarily of classifying the handedness of the

galaxies. Clever algorithms within the “Zoo” serve the same galaxy to multiple users as a reference benchmark and to check up on players to see how accurate they are.

The results from the first year’s aggregated classification of galaxies by the public proved to be just as accurate as that done by astronomers. More than 50 million classifications of a million galaxies were done by the public in the first year, and the claim about right/left handed preference was ultimately refuted. Meanwhile, Hanny Van Arkel, a schoolteacher in Holland, found a galaxy that is now the bluest known galaxy in the universe. It has come under intense scrutiny by major telescopes, including the Very Large Array (VLA) radio telescope, and will soon be scrutinized by the Hubble Space Telescope.

DEMOCRATIZING ACCESS VIA NEW BROWSING TECHNOLOGIES

The time needed to acquire data from any astronomical object increases at least as quickly as the square of the distance to that object, so any service that can accumulate custom ensembles of already captured images and data effectively brings the night sky closer. The use of archived online data stored in a “data cloud” is facilitated by new software tools, such as Microsoft’s WorldWide Telescope (WWT), which provide intuitive access to images of the night sky that have taken astronomers thousands and thousands of hours of telescope time to acquire.

Using WWT (shown in Figure 1 on the next page), anyone can pan and zoom around the sky, at wavelengths from X-ray through radio, and anyone can navigate through a three-dimensional model of the Universe constructed from real observations, just to see what’s there. Anyone can notice an unusual correspondence between features at multiple wavelengths at some position in the sky and click right through to all the published journal articles that discuss that position. Anyone can hook up a telescope to the computer running WWT and overlay live, new images on top of online images of the same piece of sky at virtually any wavelength. Anyone can be guided in their explorations via narrated “tours” produced by WWT users. As more and more tours are produced, WWT will become a true “sky browser,” with the sky as the substrate for conversations about the universe. Explorers will navigate along paths that intersect at objects of common interest, linking ideas and individuals. Hopping from tour to tour will be like surfing from Web page to Web page now.

But the power of WWT goes far beyond its standalone ability. It is, and will continue to be, part of an ecosystem of online astronomy that will speed the progress of both “citizen” and “professional” science in the coming years.



FIGURE 1.

WorldWide Telescope view of the 30 Doradus region near the Large Magellanic Cloud.

Image courtesy of the National Optical Astronomy Observatory/National Science Foundation.

Microsoft, through WWT, and Google, through Google Sky, have both created API (application programming interface) environments that allow the sky-browsing software to function inside a Web page. These APIs facilitate the creation of everything from educational environments for children to “citizen science” sites and data distribution sites for professional astronomical surveys.

Tools such as Galaxy Zoo are now easy to implement, thanks to APIs. So it now falls to the astronomical and educational communities to capitalize on the public’s willingness to help navigate the increasing influx of data. High-school students can now use satellite data that no one has yet analyzed to make real discoveries about the Universe, rather than just sliding blocks down inclined planes in their physics class. Amateur astronomers can gather data on demand to fill in missing information that students, professionals, and other astronomers ask for online. The collaborative and educational possibilities are truly limitless.

The role of WWT and tools like it in the professional astronomy community will

also continue to expand. WWT in particular has already become a better way to access all-sky surveys than any extant professional tool. WWT, as part of international “virtual observatory” efforts, is being seamlessly linked to quantitative and research tools that astronomers are accustomed to, in order to provide a beautiful contextual viewer for information that is usually served only piecemeal. And it has already begun to restore the kinds of holistic views of data that astronomers were used to before the Digital Age chopped up the sky into so many small pieces and incompatible formats.

GROWING COMPUTATIONAL POWER

In 10 years, multi-core processors will enhance commodity computing power two to three orders of magnitude beyond today’s computers. How will all this computing power help to address the data deluge? Faster computers and increased storage and bandwidth will of course enable our contemporary approaches to scale to larger datasets. In addition, fully new ways of handling and analyzing data will be enabled. For example, computer vision techniques are already surfacing in consumer digital cameras with face detection and recognition as common features.

More computational power will allow us to triage and potentially identify unique objects, events, and data outliers as soon as they are detected and route them to citizen-scientist networks for confirmation. Engagement of citizen scientists in the alerting network for this “last leg” of detection can be optimized through better-designed interfaces that can transform work into play. Interfaces could potentially connect human confirmation of objects with global networks of games and simulations where real-time data is broadly distributed and integrated into real-time massive multiplayer games that seamlessly integrate the correct identification of the objects into the games’ success metrics. Such games could give kids the opportunity to raise their social stature among game-playing peers while making a meaningful contribution to science.

VISUALIZATION AND ANALYSIS FOR THE FUTURE

WWT offers a glimpse of the future. As the diversity and scale of collected data expand, software will have to become more sophisticated in terms of how it accesses data, while simultaneously growing more intuitive, customizable, and compatible.

The way to improve tools like WWT will likely be linked to the larger challenge of how to improve the way visualization and data analysis tools can be used together in all fields—not just in astronomy.

Visualization and analysis challenges are more common across scientific fields than they are different. Imagine, for example, an astronomer and a climate scientist working in parallel. Both want to study the properties of physical systems as observed within a spherical coordinate system. Both want to move seamlessly back and forth between, for example, spectral line observations of some sources at some specific positions on a sphere (e.g., to study the composition of a stellar atmosphere or the CO₂ in the Earth's atmosphere), the context for those positions on the sphere, and journal articles and online discussions about these phenomena.

Today, even within a discipline, scientists are often faced with many choices of how to accomplish the same subtask in analysis, but no package does all the subtasks the way they would prefer. What the future holds is the potential for scientists, or data specialists working with scientists, to design their own software by linking componentized, modular applications on demand. So, for example, the astronomer and the climate scientist could both use some generalized version of WWT as part of a separate, customized system that would link to their favorite discipline- or scientist-specific packages for tasks such as spectral-line analysis.

CONCLUSION

The question linking the three topics we have discussed here is, “How can we design new tools to enhance discovery in the data deluge to come in astronomy?” The answer seems to revolve around improved *linkage* between and among existing *resources*—including citizen scientists willing to help analyze data; accessible image browsers such as WWT; and more customized visualization tools that are mashed up from common components. This approach, which seeks to more seamlessly connect (and reuse) diverse components, will likely be common to many fields of science—not just astronomy—in the coming decade.



Instrumenting the Earth: Next-Generation Sensor Networks and Environmental Science

MICHAEL LEHNING
NICHOLAS DAWES
MATHIAS BAVAY

WSL Institute for
Snow and Avalanche
Research SLF

MARC PARLANGE
École Polytechnique
Fédérale de Lausanne

SUMAN NATH
FENG ZHAO
Microsoft Research

INCREASING ENVIRONMENTAL CHALLENGES WORLDWIDE and a growing awareness of global climate change indicate an urgent need for environmental scientists to conduct science in a new and better way. Existing large-scale environmental monitoring systems, with their coarse spatiotemporal resolution, are not only expensive, but they are incapable of revealing the complex interactions between atmospheric and land surface components with enough precision to generate accurate environmental system models.

This is especially the case in mountainous regions with highly complex surfaces—the source of much of the world’s fresh water and weather patterns. The amount of data required to understand and model these interactions is so massive (terabytes, and increasing) that no off-the-shelf solution allows scientists to easily manage and analyze it. This has led to rapidly growing global collaboration among environmental scientists and computer scientists to approach these problems systematically and to develop sensing and database solutions that will enable environmental scientists to conduct their next-generation experiments.

NEXT-GENERATION ENVIRONMENTAL SCIENCE

The next generation of environmental science, as shown in Figure 1, is motivated by the following observations by the atmospheric science community: First, the most prominent challenge

in weather and climate prediction is represented by land-atmosphere interaction processes. Second, the average effect of a patchy surface on the atmosphere can be very different from an effect that is calculated by averaging a particular surface property such as temperature or moisture [1-3]—particularly in the mountains, where surface variability is typically very high.

Figure 2 shows an example of this—a highly complex mountain surface with bare rocks, debris-covered permafrost, patchy snow cover, sparse trees, and shallow and deep soils with varying vegetation. All of these surface features can occur within a single kilometer—a resolution that is typically not reached by weather forecast models of even the latest generation. Existing models of weather prediction and climate change still operate using a grid resolution, which is far too coarse (multiple kilometers) to explicitly and correctly map the surface heterogeneity in the mountains (and elsewhere). This can lead to severe errors in understanding and prediction.

In next-generation environmental science, data resolution will be addressed using densely deployed (typically wireless) sensor networks. Recent developments in wireless sensing have made it possible to instrument and sense the physical world with high resolution and fidelity over an extended period of time. Wireless connections enable reliable collection of data from remote sensors to send to laboratories for processing, analyzing, and archiving. Such high-resolution sensing enables scientists to understand more precisely the variability and dynamics of environmental parameters. Wireless sensing also provides scientists with safe and convenient visibility of *in situ* sensor deploy-



FIGURE 1. A typical data source context for next-generation environmental science, with a heterogeneous sensor deployment that includes (1) mobile stations, (2) high-resolution conventional weather stations, (3) full-size snow/weather stations, (4) external weather stations, (5) satellite imagery, (6) weather radar, (7) mobile weather radar, (8) stream observations, (9) citizen-supplied observations, (10) ground LIDAR, (11) aerial LIDAR, (12) nitrogen/methane measures, (13) snow hydrology and avalanche probes, (14) seismic probes, (15) distributed optical fiber temperature sensing, (16) water quality sampling, (17) stream gauging stations, (18) rapid mass movements research, (19) runoff stations, and (20) soil research.

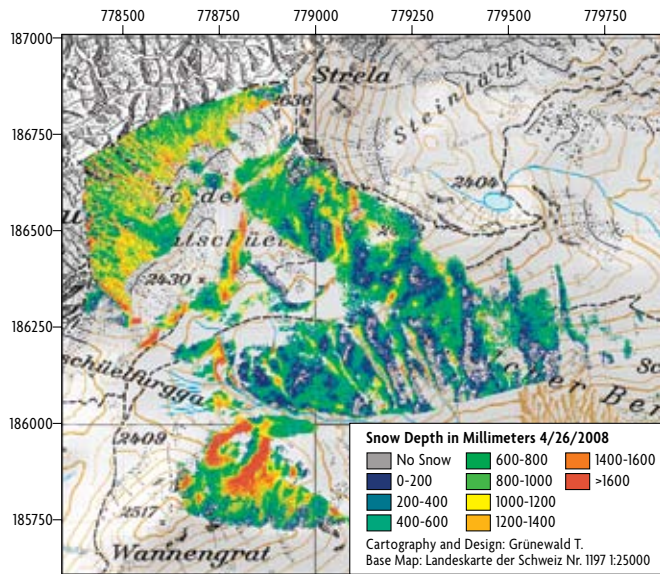


FIGURE 2.
Terrestrial laser scan for snow distribution in the Swiss Alps showing typical patchy snow cover.

ments and allows them to enable, debug, and test the deployments from the laboratory. This helps minimize site visits, which can be costly, time consuming, and even dangerous.

However, dense sensor deployments in harsh, remote environments remain challenging for several reasons. First, the whole process of sensing, computation, and communication must be extremely energy efficient so that sensors can remain operational for an extended period of time using small batteries, solar panels, or other environmental energy. Second, sensors and their communication links must be fairly robust to ensure reliable data acquisition in harsh outdoor environments. Third, invalid sensor data due to system failures or environmental impacts must be identified and treated accordingly (e.g., flagged or even filtered from the dataset). Although recent research (including the Swiss Experiment and Life Under Your Feet) partially addresses these issues, further research is needed to address them in many production systems.

MANAGING AND EXPLORING MASSIVE VOLUMES OF SENSOR DATA

High-resolution environmental sensing introduces severe data management challenges for scientists. These include reliably archiving large volumes (many terabytes) of data, sharing such data with users within access control policies, and maintaining sufficient context and provenance of sensor data using correct metadata [4].

Environmental scientists can use commercial database tools to address many of the data management and exploratory challenges associated with such a massive influx of data. For example, Microsoft's SenseWeb project [5] provides an infrastructure, including an underlying Microsoft SQL Server database, for archiving massive amounts of sensor data that might be compressed and distributed over multiple computers. SenseWeb also maintains suitable data indexes and enables efficient query processing to help users quickly explore the dataset to find features for detailed analysis [5-7]. But even with these capabilities, SenseWeb hits just the tip of the iceberg of the challenging data management tasks facing environmental scientists. Additional tools are necessary to efficiently integrate sensor data with relevant context and provide data provenance. Querying such data in a unified framework remains challenging. More research is also needed to deal with uncertain data that comes from noisy sensors and to handle the constant data flow from distributed locations.

To better understand environmental phenomena, scientists need to derive and apply various models to transform sensor data into scientific and other practical results. Database technology can help scientists to easily integrate observational data from diverse sources, possibly distributed over the Internet, with model assessments and forecasts—a procedure known as *data assimilation*. Sophisticated data mining techniques can allow scientists to easily explore spatiotemporal patterns of data (both interactively as well as in batch on archived data). Modeling techniques can provide correct and timely prediction of phenomena such as flooding events, landslides, or avalanche cycles, which can be highly useful for intervention and damage prevention, even with just a few hours of lead time. This very short-term forecasting is called *nowcasting* in meteorology.

Scientists in the Swiss Experiment project¹ have made progress in useful data assimilation and nowcasting. One case study in this project applies advanced sensors and models to forecasting alpine natural hazards [8]. A refined nowcast relies on the operational weather forecast to define the target area of a potential storm that

¹ www.swiss-experiment.ch

would affect a small-scale region (a few square kilometers) in the mountains. The operational weather forecast should allow sufficient time to install local mobile stations (such as SensorScope stations²) and remote sensing devices at the target area and to set up high-resolution hazard models. In the long term, specialized weather forecast models will be developed to allow much more precise local simulation.

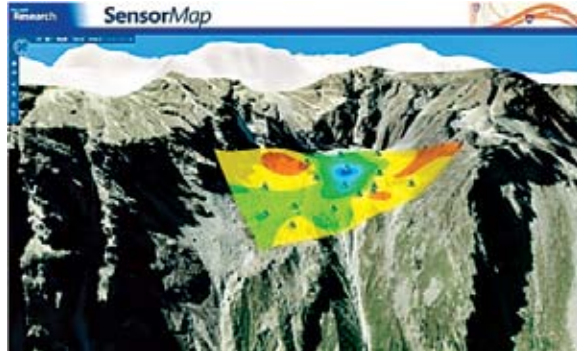


FIGURE 3. *SensorMap showing temperature distribution overlaid on 3-D mountain terrain.*

To increase the public’s environmental awareness and to support decision and policy makers, useful findings from scientific experiments must be presented and disseminated in a practical fashion. For example, SenseWeb provides a Web-based front end called SensorMap³ that presents real-time and historical environmental factors in an easy-to-understand visual interface. It overlays spatial visualizations (such as icons showing current air pollution at a location or images showing distribution of snowfalls) over a browsable geographic map, plays the visualizations of selected environmental datasets as a movie on top of a geographic map, and shows important trends in historic environmental data as well as useful summaries of real-time environmental data. (See Figure 3.) At present, such platforms support only a limited set of visualizations, and many challenges remain to be solved to support the more advanced visualizations required by diverse audiences.

WORLDWIDE ENVIRONMENTAL MONITORING

We have described the next-generation environmental monitoring system as isolated—focused on a particular region of interest such as a mountain range, ice field, or forest. This is how such environmental systems are starting to be deployed. However, we foresee far more extensive monitoring systems that can allow scientists to share data with one another and combine and correlate data from millions of

² www.swiss-experiment.ch/index.php/SensorScope:Home

³ www.sensormap.org

sensors all over the world to gain an even better understanding of global environmental patterns.

Such a global-scale sensor deployment would introduce unprecedented benefits and challenges. As sensor datasets grow larger, traditional data management techniques (such as loading data into a SQL database and then querying it) will clearly prove inadequate. To avoid moving massive amounts of data around, computations will need to be distributed and pushed as close to data sources as possible [7]. To reduce the storage and communication footprint, datasets will have to be compressed without loss of fidelity. To support data analysis with reasonable latencies, computation should preferably be done over compressed data [9]. Scientific analysis will also most likely require additional metadata, such as sensor specifications, experiment setups, data provenance, and other contextual information. Data from heterogeneous sources will have to be integrated in a unified data management and exploration framework [10].

Obviously, computer science tools can enable this next-generation environmental science only if they are actually used by domain scientists. To expedite adoption by domain scientists, such tools must be intuitive, easy to use, and robust. Moreover, they cannot be “one-size-fits-all” tools for all domains; rather, they should be domain-specific custom tools—or at least custom variants of generic tools. Developing these tools will involve identifying the important problems that domain scientists are trying to answer, analyzing the design trade-offs, and focusing on important features. While such application engineering approaches are common for non-science applications, they tend not to be a priority in science applications. This must change.

CONCLUSION

The close collaboration between environmental science and computer science is providing a new and better way to conduct scientific research through high-resolution and high-fidelity data acquisition, simplified large-scale data management, powerful data modeling and mining, and effective data sharing and visualization. In this paper, we have outlined several challenges to realizing the vision of next-generation environmental science. Some significant progress has been made in this context—such as in the Swiss Experiment and SenseWeb, in which an advanced, integrated environmental data infrastructure is being used by a variety of large environmental research projects, for environmental education, and by individual scientists. Meanwhile, dramatic progress is being made in complementary

fields such as basic sensor technology. Our expectation is that all of these advances in instrumenting the Earth will help us realize the dreams of next-generation environmental science—allowing scientists, government, and the public to better understand and live safely in their environment.

REFERENCES

- [1] M. Bavay, M. Lehning, T. Jonas, and H. Löwe, “Simulations of future snow cover and discharge in Alpine headwater catchments,” *Hydrol. Processes*, vol. 22, pp. 95–108, 2009, doi: 10.1002/hyp.7195.
- [2] M. Lehning, H. Löwe, M. Ryser, and N. Raderschall, “Inhomogeneous precipitation distribution and snow transport in steep terrain,” *Water Resour. Res.*, vol. 44, 2008, doi: 10.1029/2007WR006545.
- [3] N. Raderschall, M. Lehning, and C. Schär, “Fine scale modelling of the boundary layer wind field over steep topography,” *Water Resour. Res.*, vol. 44, 2008, doi: 10.1029/2007WR006544.
- [4] N. Dawes, A. K. Kumar, S. Michel, K. Aberer, and M. Lehning, “Sensor Metadata Management and Its Application in Collaborative Environmental Research,” presented at the 4th IEEE Int. Conf. e-Science, 2008.
- [5] A. Kansal, S. Nath, J. Liu, and F. Zhao, “SenseWeb: An Infrastructure for Shared Sensing,” *IEEE MultiMedia*, vol. 14, no. 4, pp. 8–13, Oct. 2007, doi: 10.1109/MMUL.2007.82.
- [6] Y. Ahmad and S. Nath, “COLR-Tree: Communication Efficient Spatio-Temporal Index for a Sensor Data Web Portal,” presented at the Int. Conf. Data Engineering, 2008, doi: 10.1.1.65.6941.
- [7] A. Deshpande, S. Nath, P. B. Gibbons, and S. Seshan, “Cache-and-Query for Wide Area Sensor Databases,” *Proc. 22nd ACM SIGMOD Int. Conf. Management of Data Principles of Database Systems*, 2003, doi: 10.1145/872757.872818.
- [8] M. Lehning and C. Wilhelm, “Integral Risk Management and Physical Modelling for Mountainous Natural Hazards,” in *Extreme Events in Nature and Society*, S. Albeverio, V. Jentsch, and H. Kantz, Eds. Springer, 2005.
- [9] G. Reeves, J. Liu, S. Nath, and F. Zhao, “Managing Massive Time Series Streams with MultiScale Compressed Trickle,” *Proc. 35th Int. Conf. Very Large Data Bases*, 2009.
- [10] S. Nath, J. Liu, and F. Zhao, “Challenges in Building a Portal for Sensors World-Wide,” presented at the First Workshop on World-Sensor-Web, 2006, doi: 10.1109/MPRV.2007.27.