White Paper:
Building Plausible Digital Representations of Physical Entities

Dr. Alva L. Couch and Alex Bedig, Tufts University and CUAHSI

In designing a planetary-scale data infrastructure to enable science, some notion of data trust is essential. This gives rise to many scientific questions. What is trustable data? How do trust and distrust arise? What are the foundations of trust? How can trust be fostered and enabled?

In this white paper, we will discuss the issue of data trust in hydrology, how data trust arises as a central theme in modeling hydrological objects, and what one can do to enable research by building trust-centered infrastructure. We will discuss how trust and distrust arise in practice. We will suggest a radical strategy for managing digital research data that solves perceived problems in hydrological research and has potential to solve similar problems in other geo-disciplines. This approach is based upon a trust-centered form of Bayesian reasoning about what might be called the “plausibility” of data: a measure of how much we trust it.

This differs from other white papers in the Design thread of EarthCube in discussing substructure requirements rather than overall architecture. Rather than considering how the whole EarthCube should fit together, we consider the sub-goal of creating a shared and usable digital representation of a physical entity – such as a watershed – from diverse and multitudinous data sources. While this seems on the surface to be a simple problem, modeling a physical entity from diverse data sources touches upon some subtle issues of data trust. Thus, solving the trust problems in creating a digital representation is a reasonable exemplar of how to solve trust problems in related geoscience disciplines.

In the following, we will discuss the data trust issues in constructing a digital watershed in detail, as an illustration of the issues, challenges, and potential solutions in planetary-scale scientific data sharing. Our argument is motivated by the needs of hydrologists, but the issues that arise for the digital watershed arise whenever one tries to create a representation of any complex system from diverse and heterogeneous data sources. In the context of pursuing this dream, hydrologists face discipline-specific challenges that we think will be familiar to those in other Earth sciences disciplines. Thus, we will write about the needs of hydrologists, and the reader is invited to consider how the issues that arise are similar to those in the other geo-sciences. We believe that our issues are quite similar, but for now, we will concentrate on hydrology.

The “Digital Watershed”

One dream of hydrological science is to be able to create the “digital watershed”: a complete and integrated model of a watershed based upon all available information. The digital watershed will be constructed both from data collected from multitudinous data sources and from perhaps equally multitudinous models that interpret that data and fill in unobserved details. The digital watershed is more than just a collection of data: it is a representation of “all that is known” about a hydrological entity, in the same way that EarthCube will be “all that is known” about the Earth. Thus, issues that arise in creating a digital watershed can inform the EarthCube goal of representing the Earth in a way that enables science.

At this time – with more digital hydrological data now available than ever before – the digital watershed remains a dream for deep scientific reasons that are independent of data availability. Data sources vary in quality and what one might call “plausibility – a measure of data trust. Plausibility factors include how frequently data is sampled, how frequently remote stations are checked and calibrated, how completely experimental procedures that lead to data are documented, and how measured data are vetted before publication. Building a digital watershed requires resolving the above ambiguities in data plausibility, and forming a complete picture from unavoidably incomplete data.

At present, a typical pattern is for hydrologists to take personal responsibility for the plausibility of their own data. Each individual researcher makes judgments about data source plausibility, imposes custom data filters, and completes a depiction of a watershed based upon those judgments. In creating this depiction, hydrologists tend to evaluate data based upon the publishing organization or source. USGS data is considered to be of high quality, while there are numerous individual sources of information – including University-based data collection initiatives – whose data quality is not generally agreed upon. Because hydrologists “assume ownership of their own data” by imposing custom filters and notions of plausibility, every scientific depiction of a watershed is based upon slightly different data, and data plausibility is not at present an explicitly shared scientific concept.

Thus, creating a shared concept of a digital watershed is not a simple matter of combining data, but rather, requires developing a shared scientific concept of data plausibility that does not currently exist. This requires reconciling data sources of varying quality, measurement frequency, geographic accuracy, and other factors into a complete digital picture of an entity. To be useful to scientists, the result of this must be agreed upon as a reasonable depiction of reality that is suitable for further community study. Thus, creating a digital watershed requires not only scientific mechanisms for data integration, but also, scientific governance of data collection and integration mechanisms to assure the scientific plausibility and community usefulness of the results.

We believe that the key to creating the digital watershed is to develop a persuasive and shared notion of data trust. This notion should include **a** chain of trust from data provider to scientist, and from scientist to end-user of scientific results, which determines whether data and results are plausible or useful. The consumer of integrated data needs both reasons to trust the results and some assurance that these reasons will endure over time. Wholesale data integration of “all data we have” – without explicitly handling the issue of trust – leads to a digital object that is not trustable in a scientific sense. There is a “weakest link” phenomenon that data combined naively is only as trustable as its least trustworthy source.

So, a central problem in creating the digital watershed is to define some notion of plausibility of data, where we quite intentionally avoid defining what “plausibility” exactly entails at this point in the discussion. This includes plausibility of data sources, as well as plausibility of models and data derived from those models.

A Bayesian View of Plausibility

We think that the need for a shared notion of plausibility is a crisis in planetary-scale science that requires a fundamentally different model of data trust than is used now, along with a different way of formulating the problem of fusing data sources into a digital representation of a watershed. Solving this problem requires questioning some of the fundamental tenets of how we engage in science now, and rethinking the foundations of how and why we trust hydrological data. This also requires a fundamentally new concept of modeling than we have utilized in the past in order to address the issue of trust.

One potential solution to the problem of trust is to utilize Bayesian reasoning rather than traditional statistical analysis. In the argument that follows, we are inspired by recent thinking in both High Energy Physics and Astronomy. Faced with a massive partial information problem and a problem of data trust similar to those in the Earth Sciences, astronomers have developed a Bayesian approach to scientific reasoning. Using Bayesian statistics, we can quantify the concept of "plausibility" of a data source (in the sense of the word as used by E. T. Jaynes in *Probability Theory: The Logic of Science*). Plausibility is not an absolute; it is quantified as a probability that data is plausible. In a Bayesian view, every such probability is conditional; the probability that data is plausible is always conditioned by some unknown set of environmental factors X. Further, it is assumed that we will never, ever be able to derive the exact nature of X. Thus we make a "maximum entropy assumption" that the safest scientific assumption for the nature of the unknown factor X is that which maximizes the entropy in the overall system.

One key to Bayesian reasoning is never to assume that a data point is an "outlier." In hydrological science, there are many stories of presumed outliers that seemed to be exceptional only due to missing facts. For example, discrepancies between upstream and downstream flow measurements may not have anything to do with sensor accuracy at all; there may be an unknown fact that influences the result (e.g., a new and previously unmapped stream channel). When two data sources do not agree, there is some unknown factor Y – perhaps distinct from the correctness of the data sources – that led to the discrepancy.

The Bayesian view of scientific data differs dramatically from the view promoted by classical statistics. While it might be reasonable in classical statistics to declare a discrepancy as an outlier, in the Bayesian view of an integrated data set, an apparent outlier remains – in perpetuity – evidence of some unknown set of factors Y. Such a characterization remains a hypothesis forever, and can become more or less likely or unlikely over time, based upon other accumulated evidence. Thus, the global picture changes with the addition of each data source, and the plausibility of data becomes a fluid thing, that changes with each addition of new evidence. In other words, there is no such thing as an outlier; there is only hidden information.

Hydrologists have already discovered the power of Bayesian reasoning. Keith Beven has observed that sources of hidden information exist in the task of watershed modeling, and endorses a Bayesian approach as a reasonable way to conduct technical and policy decision-making in the face of the implied uncertainty. Common sources of hidden information include "non-stationarity in the errors of estimates of inputs to a catchment system, unknown temporal in system characteristics as represented by model parameters, unknown temporal and spatial variability in controlling processes, and indecision about how some processes should be represented mathematically." (Bevin and Alcock, "Modelling everything everywhere: a new approach to decision-making for water management under uncertainty", *Freshwater Biology,* 2011).

Building a Bayesian Digital Watershed

Building a digital watershed requires a careful understanding of the ground truths of the data store, i.e., those data sources whose "plausibility" is 1.0. Ground truth is expensive in human effort, and involves intensive and ongoing data validation, e.g., regular recalibration of sensors. This also requires ongoing governance and decision making about which organizations and individuals create and curate highly accurate data sources. For hydrologists, the ground truth is often USGS data, but this is only because that is the only convenient ground truth data at this time.

Starting with a set of ground truths, a Bayesian digital watershed can be constructed using a somewhat novel form of hydrological modeling. Most past hydrological models have been intended for prediction of future data and/or filling in unobserved details of a watershed. A new form of modeling is needed, that serves to cross-validate questionable data sources with respect to mutually agreed-upon ground truths of the watershed. This modeling takes as input a set of data sources with known plausibilities, as well as a data set with unknown plausibilities, and estimates the plausibilities of the unknown sources from those with known plausibilities via a method similar to that in the example described below. Constructing these new models is a matter of re-crafting some commonly agreed-upon hydrological models as cross-validation models.

One strength of Bayesian reasoning in building the digital watershed is that it allows the scientist not just to codify what is unknown, but also, what is unknowable, in the sense that there is no practical mechanism for discovering it. For example, every hydrological model is only plausible in certain environmental circumstances. Try as we might, specifying the exact circumstances precisely is impractical; the whole set of circumstances that make a model plausible is in essence unknowable. In classical statistics, we would make an assumption about plausibility, while in Bayesian statistics, we instead give a name to the circumstances whose exact nature remains unknown.

While the implementation of a Bayesian reasoning model is complex, the underlying mathematics are relatively simple to describe on the surface. As an oversimplified example, suppose we have a model M that is plausible in some unknown environmental circumstances X, and that our actual environmental circumstances are represented as Y. Suppose that for simplicity that we express plausibility as a probability (which may or may not be the best way to express plausibility; further study is needed). Then the ideal plausibility of M given X (Prob(M/X)) is 1.0, but the realistic plausibility of M is a conditional probability Prob(M/Y) based upon actual circumstances. Given new data D, a properly designed model allows us to compute the plausibility of the data according to the model M as Prob(D/M), and then the Bayesian expression Prob(M/Y) Prob(D/M) is the probability Prob(D/Y), which represents a model-independent notion of plausibility of data D in environment Y. What makes Bayesian reasoning particularly suitable for this kind of analysis is that we can estimate Prob(M/Y) by examining the behavior of M on data known to be plausible, and thus estimate Prob(D/Y). Overlapping estimations of Prob(D/Y) based upon different models thus lead to a model-independent estimate of the plausibility of D based upon all available information.

While the above example is very much over-simplified for brevity, its point is that one can judge the plausibility of any new data based upon data that is already present and whose plausibility has itself been judged. Conversely, one can judge the plausibility of a model based upon its relationship to existing plausible data. In this way, every data set (or model) in the digital watershed can be assigned a plausibility that determines how much that this data set (or model) should be trusted according to all available information. At global scale, labeling each data set in this fashion is a supercomputing task, and labels may change as the state of knowledge changes.

How the Digital Watershed Enables Science

Hydrologists will be able to interact with the proposed digital watershed in a fundamentally different way than the way that they currently interact with data repositories. Through Bayesian reasoning, the digital watershed can establish its own ideas of plausibility and outliers, independent of those of the scientist. Scientists will interact with it not by filtering data (as they do now), but by giving the digital watershed new data and/or models to incorporate. Models can be tested for plausibility at a watershed scale, by checking them against existing data, while new data is checked against plausible models. The scientist measuring data will “throw it into the digital watershed” and immediately obtain some measure of its plausibility in relation to plausible data sources and models. The scientist testing a model can “throw it into the EarthCube” and determine every watershed on the Earth where the model is plausible. Supercomputing will enable checking the global validity of an idea as easily as one could check local validity before.

Next Steps and Challenges

Enabling this kind of science takes several scientific steps, including addressing issues in:

1. Governance:
    a. Developing a shared notion of data plausibility.
    b. Deciding upon the data sources that will be considered as highest quality.
    c. Deciding upon the commonly accepted models that will be used for cross-validation.
2. Modeling:
    a. Building models of plausible relationships between measurements in a local vicinity.
    b. Reconciling plausibility relationships that arise from overlapping local vicinities.
    c. Reconciling relationships between planetary-scale and local-scale data (e.g., comparing satellite and ground-based data).
3. Infrastructure:
    a. Establishing a language for describing model validity factors.
    b. Embedding plausibility metadata into data sources.
    c. Enabling planetary-scale data and model validation via super-computing.

The grand challenges of this technique seem more related to governance rather than technical implementation. The scientific community has yet to incorporate this kind of global thinking into its workflows. The digital watershed – if constructed properly – will become a scientific object with a life of its own, independent from the researchers who contribute to it. Researchers require assurance that the results of interacting with shared digital objects (such as the digital watershed) will be credible, attributable, and publishable. The greatest challenge seems to be to build bridges between traditional Earth sciences workflows and the new paradigms represented by this kind of work.

Conclusions

We often think of science as "producing archival results" in the form of publications and what many have called "durable knowledge." In an age of supercomputing and planetary-scale data integration, the concept of durability of knowledge is evolving. It is hypotheses that endure – or not – and the publications are merely a byproduct. These hypotheses have their primary lives in the data repositories, and not in the publications. The so-called "durable knowledge" consists of hypotheses that stand the test of time and scale.

The existence of persistent digital descriptions of Earth entities will eventually change the way science is done. Engaging in science will be a matter of "throwing a hypothesis into the EarthCube" and testing its validity. Crucial to this process – however – is the concept of trust in the EarthCube itself. The concepts in this white paper are the beginning or foundation for such a concept of trust, which will both enable a new kind of computational science and carefully define its notions of scientific truth.