J. Jpn. Soc. Soil Phys.
土壌の物理性
No. 134, p.51 ～ 59 (2016)

資　料
Miscellaneous

Improving the impact of soil science by reconsidering the interactions of data, models, and decisions — An ongoing discussion during the 2016 Darcy lecture tour

Ty P.A. FERRÉ [1]

Abstract: How can we improve the acceptance and inclusion of science in public debate and decision making? In this article, I argue that scientists have to do a better job of understanding what drives people to make decisions. By listening to stakeholders concerns, we can improve our science and its impact.
Key Words : data, models, decision making

1. Introduction

The following is a discussion born out of the 2016 Darcy Lecture. This lecture series, sponsored by the National Ground Water Association's Research and Education Fund (NGWREF), allows one hydrogeologist per year to travel around the world to meet with the hydrogeologic community. The lecturer can choose any topic, or set of topics, to present. Hosts can invite the lecturer, who then constructs the final lecture schedule.

I was honored to have been asked to provide the 2016 Darcy Lecture. My research has touched on several areas: hydrogeophysics; vadose zone hydrology; and optimal monitoring network design. After some thought, I decided that I wanted to focus my lectures on a topic that is relatively new to our research group: how we can improve the impact of science in water-related decision making. This brings together many elements of my research. But, more importantly, I think that it is a topic that is well suited to promote conversation and exchange among academic researchers, practicing hydrogeologists, and members of the public who need help making difficult water-related decisions. With this in mind, I laid out an ambitious schedule of over 120 talks, including leading research universities, education-focused universities and colleges, professional groups, stakeholder groups, and public schools. The goal of the lecture series was to propose some changes that we, as scientists, can make in the way that we approach our work and communicate our findings. Simultaneously, I hoped to provide non-scientists with a useful basis for communicating with scientists to ensure that science serves their needs. The combined goal is to ensure that we do our part to ensure that science will be included in public debate.

This is an updated version of a paper published in the proceedings of the recent meeting of the Japanese Association of Hydrologeologists. At the time of writing, I have completed 112 talks in North America, Western Europe, South America, South Africa, Australian, New Zealand, Israel and Japan (Nagasaki, Tokyo, Mie, and Kyoto). In this paper, I focus on bringing together the ideas that have developed through these talks and discussions that followed with special emphasis on topics related to soil science. I hope that this article will encourage readers to contact me to continue discussions on the topics presented. Those who are interested can find more detailed information about my talk, my contact information, a list of previous questions, and related references on my blog: https://darcylecture2016.wordpress.com/.

2. My background

I consider myself to be, at least in part, a soil physicist. As a graduate student, I was drawn to soil physics because of the nonlinearities that lie at its heart. The dependence of hydraulic conductivity on water content, and water content on water potential, lead to surprising, unexpected outcomes — even if you already have a solid background in hydrogeology as it applies to the saturated zone.

My specific research interest was in the area of nondestructive measurement of water content, specifically using time domain reflectometry (TDR). Through my PhD research I learned that many indirect measurement meth-

[1] Department of Hydrology and Water Resources, University of Arizona, Tucson, AZ 85721-0011, USA.
2016 年 9 月 26 日受稿　2016 年 11 月 18 日受理

ods, like TDR, include their own nonlinearities that can lead to surprising results. In particular, the spatial distribution of water in the volume surrounding a TDR probe can have important impacts on what the probe measures. As a result, a more complete and accurate interpretation of water content measurements with TDR should often include consideration of the sub-sample-volume water content distribution. But, how can we do this if we don't have any way to measure this distribution? The answer, is to model the processes of interest, use the model to propose the water content distributions, use these to model what TDR would measure given this proposed distribution, and then use the TDR measurements to test the result. If the TDR measurement does not agree with the predicted value, then the underlying soil physical model must be changed. This was the genesis of an approach that has come to be known as coupled hydrogeophysical analysis. The major lesson that we took from these investigations is that, once we couple many complex systems, we must consider that changes to any of these systems (soil hydraulic property dependencies, unsaturated flow, TDR spatial sensitivity) will affect the fit between the model and the data. Conversely, changes to these systems can compensate for one another, meaning that a good fit between data and model results does not prove that all of the components of our model are correct.

As I continued in my research career, I expanded my research to other geophysical methods (gravity, neutron probes, electrical resistivity, electromagnetic methods, temperature profiling, and nuclear magnetic resonance) and to processes beyond the root zone. But, the basic concept of coupled measurement/model analysis has been consistent. More recently, I have started thinking about how our analyses, as soil scientists, hydrologists, or scientists in general, are used by decision makers. I have become especially interested in examining why science is, too often, ignored in public discussions and decision making processes. My conclusion is that we, as scientists, must do a better job of considering another type of coupled analysis — how decision making and scientific investigations influence one another. The purpose of my Darcy Lecture has been to discuss how science can do a better job of providing the kinds of analyses that decision makers can and will use, while still maintaining the objectivity that is the foundation of science.

3. Science and decision making

Humans are decision-making machines. We make thousands of decisions, large and small, every day. But, we often fail to consider how we make decisions, leaving us open to common failures in decision making. Recent research has led to greatly improved insights into the processes of decision making. In the popular literature, the range of these findings are well described in two books: Incognito, The Secret Lives of the Brain (Eagleman, 2012) and Thinking Fast and Slow (Kahneman, 2011). These authors point to the different impacts of detailed versus structural uncertainty in decision making. Specifically, Eagleman suggests that our brains have evolved to make decisions under considerable detailed uncertainty because our senses are imperfect sensors of the world around us. Kahneman suggests that we resort to shortcuts and unconscious biases when confronted with more fundamental, structural uncertainties. That is, when we don't fully understand a system, we make simplifications to systems that we can understand, and then we fall into habitual approaches to making decisions in ways that 'feel' justifiable.

It behooves us, as scientists, to recognize that the information and insights that we provide will be viewed in the context of flawed decision making processes. In particular, we must consider that structural uncertainty, which we often face when representing natural systems, will be treated with some degree of illogical thinking. Further, that we all tend to hold to the stories that we have constructed to make sense of complex systems and we resist arguments to change our ways of thinking. Another popular book, Nudge, Improving Decisions about Health, Wealth, and Happiness (Thaler and Sunstein, 2009) is relevant to this point. This work has been applied widely in advertising and is based on the following premise: people are more likely to change their opinions (for example, to be informed by scientific results) when you make it easy for them to do so. My Darcy Lecture has evolved into an examination of ways that scientists can conduct their investigations in ways that are more likely to nudge decision makers to more scientifically sound decisions while maintaining the objectivity and rigor that marks science as a unique way of seeing the world.

In working with students of all ages, from pre-school to graduate school, I have also come to believe that humans are also scientists by nature. That is, we naturally employ the scientific process of proposing a hypothesis, seeking data to challenge the hypothesis, and revising the hypothesis based on observations. We seek underlying and general truths and find security in the knowledge of how the world works.

Major difficulties arise when we must make decisions for problems that are not addressed easily by science. In general, these problems are marked by a high degree of complexity in the underlying systems: psychological problems and the function of the human brain; economic problems and the function of groups of individuals; climatological problems and the interactions of many entangled

systems; and geologic problems that rely on processes that occur at multiple temporal and spatial scales that are often difficult or impossible to observe. Despite how much time and effort humans have devoted to understanding soil, the interactions of physics, chemistry, and biology in soil make this one of the most complex systems that we can study. How can we apply simple, fundamental scientific understanding to a question as complex as the cycling of nutrients in soil under different proposed management schemes? How can we describe our scientific findings in a way that correctly identifies what we know and what we do not, yet, know so that decision makers can rely on the results and so that the public will understand the need for continued scientific study?

4. Pure and applied science

Conversations that I have had over this year have led me to believe that there are two kinds of science and that we too often confuse the two. I will use the term 'pure' science to refer to the search for general truths. I will call the application of the findings of pure science to complex systems as 'applied science'. Generally, when we conduct pure science, we are careful to control as many variables and conditions as possible to allow us to isolate the effects of the processes of interest to us. Some realms of science are particularly able to do this, ranging from chemistry, to physics, to astronomy. Others sciences do not allow for controlled experiments on ethical grounds: economics; psychology; environmental sciences. Most sciences simply involve too many coupled systems to admit highly controlled experiments — soil science and hydrology fall in this category.

My claim that soil science is not a 'pure' science is not intended to say that we cannot conduct controlled experiments, such as infiltration tests, or variable irrigation/fertilization tests. But, we must always recognize that the interpretations that we use to interpret these tests may not be appropriate if the conditions do not match those for which the analyses were developed. The clearest example is the effect of heterogeneity on the interpretation of infiltration tests. Most of the analyses that we use to interpret the test results are based on an assumption of homogeneity and then return one 'average' property value. But, we know that if we were to relocate the test, often even by a meter or less, we may get very different results. Given this uncertainty, how can we build scientific models? As I see it, there are three approaches: detailed modeling with distributed measurement; upscaling with large-scale measurements; and correlative modeling using all measurements available.

It is natural to assume that the problem of heterogeneity is reduced if we can measure at a scale that is smaller than the spatial scale of the heterogeneity. This suggests that if we could collect many measurements, we could develop small scale, pure-science-based models for each location and link them together to simulate a larger system. In some cases this is true, and the development of smaller, lower power, more easily networked sensors is supporting this approach. But, in other cases, the required scale of measurement to support a pure-science representation is simply impractically small. Consider, for example, discussion related to the appropriate scale at which to consider reactions in porous media, or the transport of nanoparticles through soil, or the function of microbial communities in soils and on plant roots.

If we conclude that it is impossible to measure at a scale at which conditions are sufficiently homogeneous to apply pure-science models, then we must consider some version of upscaling. In general, all of these approaches can be distilled into some version of the following. We will assume that the pure-science models that we have formulated for homogeneous conditions apply for heterogeneous media if we can identify the appropriate averaging to provide 'effective' parameter values. Of course, this cannot be expected to apply when considering processes that involve many coupled systems (consider microbial growth that depends on solute flux that depends on water flow in a variably saturated medium averaged even on a 1 cm^3 scale). But, even for those systems that may allow for upscaling, it is naïve to believe that the manner in which our measurement methods and their associated interpretations average properties or states is the correct average for the process of interest to us. Therefore, even our most strictly pure-science-based models must be seen as some approximation of a real system.

Given the limitations of pure-science based models for real systems, there is increasing interest in the development of correlative modeling. This includes approaches such as statistical inference, neural network modeling, and many approaches collectively referred to as big data or machine learning. At the risk of oversimplifying, these approaches abandon the development of pure-science models, assuming that all of the limitations stated above make them impractical for real, complex systems. Rather, these approaches search for mathematical or statistical relationships between data and outcomes of interest. Generally, many simple models are proposed and selected primarily on the basis of their ability to predict past observations. The danger inherent in these approaches is that they can produce models that have no basis in truth — they can mistake correlation for causation. This can lead to incorrect projections and can lead to incorrect conclusions about underlying scientific concepts.

Based on the preceding discussion, it would be easy to conclude that science is doomed. If we cannot rely on any of the three general approaches to building models of complex systems, then what can we do? How can we hope to make a contribution to the important decisions that society must make and that should be informed by science? What I have learned over the course of this year is that there are two important elements that we, as scientists, must remember: the power of storytelling; and the value of developing multiple competing hypotheses.

5. Harnessing the power of storytelling

Every culture has developed a form of storytelling. This is a fundamental approach that humans use to collect disparate facts, put them into a consistent context, and weave a coherent explanation of the world. While this is a useful, even vital, ability, it also has potential limitations. Specifically, most of us are highly resistant to changing our internal narrative once it is formed. We fall in love with the story in our head and will require compelling evidence to change it.

Consider a scientist who is presenting their results based on their current best representation of a complex system. If their results strongly suggest a change in the way that the public should understand the system, and if this change in conceptualization requires a change in behavior that is difficult or costly, the public will look for reasons to discredit the science. On the other hand, if a scientist first listens to the concerns of a stakeholder group and then translates their story into a scientific model of the system, they have an opportunity to demonstrate weaknesses of that model (if they exist) and to introduce more scientifically supportable concepts. This requires a change in the way that we do science — it requires that we actually embrace one of the basic tenets of science. That is, we must view all of our work as an effort to challenge hypotheses. This can only be done by developing multiple competing hypotheses. Unfortunately, as I stated above, this runs counter to our natural tendency to prefer one unchanging internal narrative.

Through discussions across many diverse groups, the theme of the power of storytelling has emerged as a potentially strong model for explaining science. In particular, I have come to believe that we can view many multi-party water issues as multiple groups that believe different stories based on a common set of facts. There are times that a group's concerns run counter to basic science. But, more often, they are based on incomplete or inaccurate understanding of science. This is to be expected when systems are complex, involving many interacting processes. In fact, this is one reason that we, as scientists, build models — to integrate knowledge in ways that are difficult to integrate 'in our heads'.

Following on the concept of nudging, I think that it could be very useful for us to view the model building process as an attempt to construct as many scientifically plausible stories about a system as possible. In this way, we should act like detectives — simultaneously using all available information to propose many different possible suspects, then building a story that would implicate each suspect, then finding the key pieces of evidence to assess the likely guilt or innocence of each suspect. Our current approach would not be acceptable in a police investigator — decide upon the best representation of a system that we can currently conceive (choose a suspect) and then examine how we have to modify parameter values (massage the circumstantial evidence) to fit the model to the data (to prove their guilt).

In an applied context, this proposed approach would require us to make every attempt to construct plausible models that could manifest stakeholders' specific concerns. Based on the discussion of the benefits and weaknesses of our three general approaches to modeling, I would contend that we should consider all modeling approaches: detailed and spatially discrete; upscaled; and correlative. For most applied cases, the ultimate 'truth' of the models is not the primary aim. Rather, we are interested in using models to either: a) draw a consensus from multiple perspectives to identify reliable courses of action; or b) identify important differences in projections based on different viewpoints, thereby prioritizing further scientific efforts.

The proposed modeling effort would thus require more consideration of multiple conceptualizations, perhaps requiring simpler models that are less rigorously calibrated to data. The approach will require that we spend more time in the creative elements of model building, and less on model calibration. It may require that we have a more balanced view of the value of pure-science and applied science models. It will also require that we develop tools, like those that have been developed to assist with parameter estimation and uncertainty analysis, to make model construction more efficient and more universal. Ultimately, we need to view an ensemble of models like a population — the success of the species lies in the ability of the group to adapt to new conditions, rather than in the ability of the currently most-fit animal to excel. As conditions change, usually gradually, there will be differential breeding success within the population. Equivalently, as we continually gather new data, we should expect to see relatively small changes in the weight that we place on individual models in our ensemble for decision support. But, at times we will be faced with major 'surprises' (Bredehoeft, 2005). At these times, we will be forced to rethink our model ensemble,

possibly adding new processes or structures. This process of model building should come to resemble the punctuated equilibrium attributed to biological systems as explained in popular literature (Gould, 1980).

The most direct benefit of this approach is that it provides testable representations that serve as proxies for stakeholders' general concerns. The more general advantage is that in forming multiple competing models we are actually following the tenets of the scientific method: generating hypotheses that can then be tested objectively, through observation, against counter hypotheses.

6. The role of science in decision support

Through conversations with a wide audience, I have come to another realization: in most cases, science doesn't answer questions, it only helps to answer questions. This may seem like a distinction without a difference, but it has important implications for the way that we choose to pursue applied scientific problems. Essentially, two things are important for us to remember. First, the scientific element of a question is often only a part, sometimes a small part, of the discussion. Second, as physical scientists, we may be able to predict outcomes of actions on physical systems, but we are not equipped to translate these outcomes into meaning for stakeholders. The latter point requires that we work with social scientists and economists to develop relations between our predictions and stakeholder utilities. This concept is shown schematically in Fig. 1. Here, a hydrologist may be able to predict drawdown beneath a stream (with uncertainty) due to pumping for a nearby proposed mine (Fig. 1a). But, these predictions must be translated to perceived value, or utility, to determine how each group views each predicted outcome (Fig. 1b). These utility functions are at once critical to understand decision making and extremely challenging to create. Entire fields of study are devoted to this process, which I cannot review here. Rather, I will point out that the key element of these curves for our purposes are that they define the relative importance of different possible predicted outcomes for different interested parties.

My students I have been working on ways to bring together the curves presented in Fig. 1 to define a new approach to science for decision support. We refer to the approach as DIRECT — the Discrimination Inference to Reduce Expected Cost Technique. DI refers to an approach to identify additional data that are most likely to reduce prediction uncertainties in ways that are useful for decision makers. RECT refers to different approaches to making decisions under uncertainty. DIRECT is based on combining the curves presented in Fig. 1, recognizing that they share a common x-axis. This allows us to form utility probability density functions, as shown in Fig. 2.

The curves shown in Fig. 2 can be used to support three very different decision styles. Simple decision making only considers the most likely perceived value that a group will have and is based on the maximum likelihood value of each curve (the peak). As scientists, we are often guilty (although unaware) of assuming that this describes the decision making process — stakeholders will logically process scientific results, accepting uncertainty, but being willing to make decisions on the current best representation of scientific understanding. There is increasing realization that decisions are more robust if they consider lower probability outcomes, rather than focusing solely on the most likely prediction. (Again, Eagleman suggests that we have evolved to integrate such 'what if' strategies into our decision making even at the level of brain function.) This is the next level of abstraction of the decision making process — quantitative risk assessment based on the

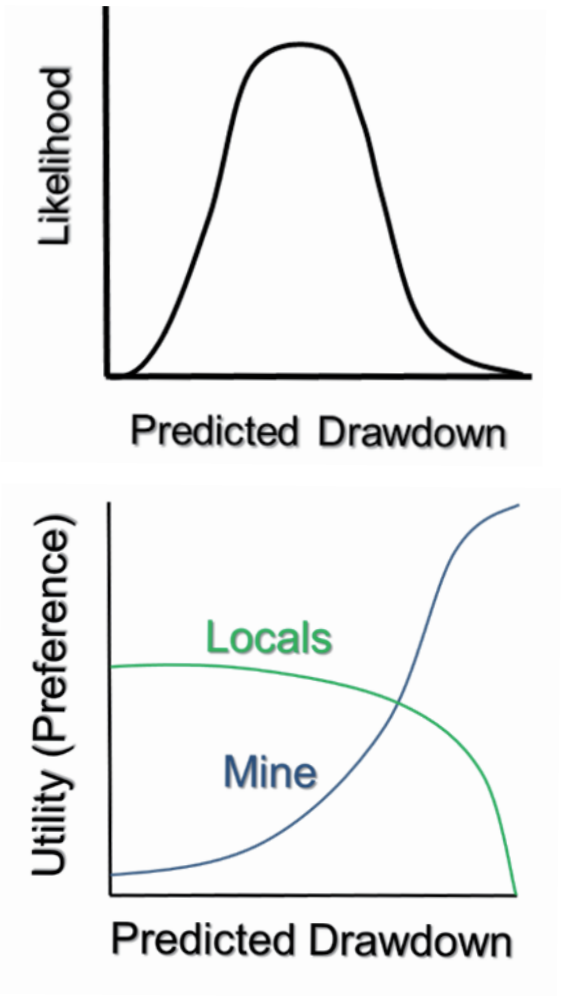

Fig. 1 a) Multi-model prediction of a drawdown of interest made by hydrogeologists; b) valuation of drawdown for two stakeholders groups (a mine and a local community) made by social scientists.

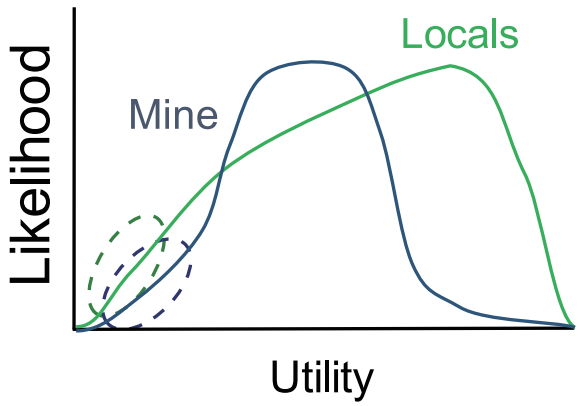

Fig. 2 a) The probability density function of utility for each stakeholder group formed by combining the two curves in Fig. 1. The dashed ellipses identify relatively low probability, but high risk outcomes that often drive risk-averse decisions.

likelihood-weighted (or expected) utility. This is also a fully logical approach and it is based on a strong assumption that the uncertainties that we have defined represent the actual uncertainties that are most relevant for decision makers. But, much of Kahneman's work focuses on the idea that humans rarely make logical decisions when faced with structural uncertainty. This leads to the third decision style, which can be described as threshold decisions or risk averse decision making. Regardless of our desires to represent decision making as scientifically informed and logical, this may be the most common approach in practice. One key element of risk averse decision making is that, whether consciously or not, decisions are often driven by relatively low probability outcomes that carry very high risks (low utilities) if they occur. These outcomes are highlighted by dashed ovals in Fig. 2. With this in mind, I would contend that science can have greater impact on decision making if it is addressed at testing these outcomes objectively and openly. That is, we will be more effective if we take a nudge approach to decision support by trying to develop models that predict outcomes of concern to decision makers and then make efforts to test these associated models against alternative explanations that lead to different outcomes.

7. DIRECT as an approach to implementing nudging

At base, DIRECT attempts to reverse our usual approach to scientific analysis. As natural scientists, we tend to start with efforts to understand a system in the most general terms. Then we condense our understanding into a model representation of the system. Finally, we use our models to make predictions that can form the basis of decision making. We are proposing that efficient use of science should start with consideration of the decision being supported, including the outcomes of concern that are driving the decision making process. Models should then be constructed to describe as wide a range of plausible descriptions of the system as possible, with special efforts taken to identify models that could predict outcomes of concern to each stakeholder group. As discussed above, these models should encompass different approaches to modeling (distributed and discrete, upscaled, and correlative). They should also consider all of the simplifying assumptions that we make in developing each of these models. In particular, we should examine those assumptions that are hardest to defend and that, in our professional judgement, have the greatest likelihood to affect predictions of interest. These models, which can be viewed as competing hypotheses, can be used to identify discriminatory data — data that are best able to test one subset of models against other competing descriptions of the system (models). In practice, DIRECT achieves this as follows:

1. A diverse ensemble of models is proposed. These models must be plausible based on current scientific knowledge and provide an acceptably good fit to existing data. But, only limited efforts are made to fit the models to the data through extensive calibration.
2. The models are used to make predictions of interest. The result (as in Fig. 1a) provides a measure of the likelihood of predictions of concern. Then, the ensemble is divided into contrasting groups, generally based on whether they make predictions of concern to a given stakeholder group (see Fig. 1b). Models that produce very low utility outcomes with sufficiently high probability (the predictions within the dashed ellipses on Fig. 2) are identified as models of concern.
3. A future measurement is proposed. The expected measured value is predicted using all of the models. A discriminatory index (Kikuchi et al., 2015) is defined based on how different the observation is predicted to be between models of concern and all other models in the ensemble. This is repeated for many possible measurements to identify measurements that are most likely to improve our understanding of the likelihoods of outcomes of concern by testing their associated models.
4. Combinations of these promising observations are examined to define measurement sets that minimize redundant information.
5. The data are collected and the likelihoods of all models in the ensemble are revised. The likelihood of the outcomes of concern are updated and the decision maker determines if further investigations are necessary.

8. A soils-related example

The case study that I highlighted in my Darcy talk was practical and related to contaminant hydrogeology — how can we best advise a client who may have to design, build, and operate a groundwater treatment facility? But, the ideas put forward in DIRECT are equally applicable to a range of practical and more purely scientific questions. Here, I will discuss briefly a practical and a scientific example that relate more directly to soil science.

As an illustrative example, we can consider groundwater protection related to the flushing of N through the root zone of an agricultural field. To understand and manage the problem, we need to develop a model. We may propose an upper boundary condition — periodic drip irrigation with infrequent high intensity rainfall, a crop based model of ET, and diurnally fluctuating evaporation. Then we propose soil hydraulic properties — single porosity, two layers (each homogeneous and isotropic), no preferential flow, the Kosugi hydraulic function. Then chemical properties — no degradation or transformation, linear reversible sorption, no internal sources or sinks. Finally, we propose root conditions – root water uptake can be distributed functionally by depth, no root growth, and roots do not affect soil hydraulic or chemical properties.

Any one of these many assumptions that underlie our model could be the subject of one or more PhD dissertations. That is, we cannot make a clear and confident decision regarding whether the assumptions that we made are correct for the system under study. Our usual approach, driven by the need to make progress in a timely fashion, is to make a set of assumptions (build a model) and move forward (calibrate the model to our data). We will generally only change one or more of our assumptions if we cannot make our model fit our data 'acceptably well' using 'reasonable' parameter values. That is, unless we are forced to do so by clear and compelling data, we will not revisit our initial assumptions. Unfortunately, this is not really a scientific approach. Essentially, we are working very hard to confirm our hypothesis (encoded as a model). As scientists, of course, our job is to try to test our hypotheses — the opposite of our usual practice. The discussions that I have had as part of my Darcy Lecture focus on how we can establish an approach to modeling that takes advantage of continually testing our assumptions while still allowing for use in solving practical problems.

The key step to our proposed approach is to force ourselves to formulate multiple competing hypotheses (models). To do this, we need to identify those assumptions that we cannot defend solidly and that, if we were to change them, may still lead to plausible models given the data that we have. This list of assumptions can guide us to develop a family, or tree, of related models that cover our fundamental, structural uncertainty. In practical cases, and when we are working as part of a multi-disciplinary team that must coordinate our investigations, we need to prioritize consideration of those assumptions that, if changed, may lead to outcomes of concern. That is, in practical cases, we should focus on building models that, if they are true, would have major impacts on decision making. If we believe that we have conditions that are best suited to one of the three general approaches to modeling described above, we can focus on building multiple models of that type. But, more generally, we should be open to considering multiple modeling approaches as well.

For scientific studies, we need to focus on the models that make predictions that would have larger or more important impacts on other scientific studies. (For example, fish biologists may be interested in stream temperature at specific locations and times. To feed into their work, we would want to focus on multiple models that may predict different groundwater fluxes through time, even if this focus reduces our ability to answer other questions about the hydrologic system.) In this way, 'pure' and 'applied' problems can be seen as highly comparable. The only real difference is the basis used for determining the outcomes of greatest interest.

Once we have formed a diverse model ensemble, we can subdivide our models into two or more groups. This can be based on a prediction or predictions of interest. For our example, this may be some threshold or regulatory limit concentration of N reaching the water table. Or, for purely scientific studies, subdivision can be based on underlying hypotheses. For example, we may want to know if it is possible to infer whether sorption is reversible given all other uncertainties that we have about our system. As described by Kikuchi et al. (2015), discriminatory data are those that are predicted to be most distinctly different based on the imposed division of the models. That is, we can formulate many models, differing in their treatment of sorption and in many other aspects, and predict what we expect to measure at any proposed measurement location and time. If the predicted values that we would observe are for the models in different model divisions, then the observation is informative and should be collected. Measurements that do not cluster in accordance with the model subdivisions are not likely to test one subset of models against the other set or sets: those data are non-discriminatory and should not be prioritized. For the example problem described, a practical investigation may be most interested in the seasonal-average N loading to an aquifer. A scientific investigation may be interested in the role of macropores in delivering N to the aquifer. It is likely that different measurements

(type, locations, spatial and temporal resolutions) will be optimal for these different objectives.

Our proposed approach has practical benefits for both applied and purely scientific studies. In both cases, the number of measurements that we can collect is limited. Some experiments cannot allow for the disruption of taking many samples, others are too remote to allow for extensive sampling, still others are budget- or time-limited. We propose that it is our responsibility to propose multiple competing hypotheses that are plausible given existing data. This encourages us to think more critically about our models, allows us to define data that are likely to test our models once collected, and avoids the collection of non-discriminatory data. In the end, we predict that this will cost less than our current approach and will lead to better and more useful science.

9. Challenges to implementing DIRECT

To date, the underlying ideas of DIRECT seem to resonate with both pure and applied scientist, regulators, and members of the general public. But, there are several questions that arise consistently. In general, concerns are focused on the practicalities of developing diverse model ensembles. I think that these concerns are based on the fact that DIRECT requires us to think differently about the purpose and process of model development. We are accustomed to viewing a model as the current best description of a system. Our model building culture has developed to make firm decisions about model structure, often despite considerable uncertainty. Then, we take great pains to adjust model parameters to fit our model to existing data. It can be argued that this is a clear example of confirmation bias — we have a hypothesis (our model, a suspect) and we actively seek to make it agree with our data (calibration, framing). This runs counter to the idea that scientists should act as the most ardent critics of their own ideas by seeking the data that are most able to disprove them. We contend that it is more scientifically valid, in addition to being more useful, to intentionally develop an ensemble of divergent plausible models. Thus, as suggested in Team of Rivals: The Political Genius of Abraham Lincoln (Goodwin, 2012), as Lincoln did when forming his cabinet, we would form a set of rival models that provide greater confidence in their areas of agreement and important insights in their areas of disagreement. Affecting this change in our approach to modeling will require changes in our view of the purpose of models. It will also require significant changes in our approach to model building. In particular, it will require that we develop tools that mimic currently available automated parameter estimation algorithms that are aimed at exploration of model structure. For scientists, the most difficult challenge in implementing DIRECT is to fight our natural human preference for holding a single, unambiguous internal narrative. The ability to formulate competing hypotheses has been praised highly in science — but, we still seem to prefer studies that end with a single, unambiguous conclusion. Of course, the greatest challenges lie beyond the realm of science. We are, at best, advisors to decision makers. We can do a better job of identifying scientific advances that are most likely to have a positive impact on decision making. We can also commit to focusing our science in ways that provide useful descriptions of scientific uncertainty. But, ultimately, the impact of science will depend on the willingness of decision makers to act on the best information that science can provide. In this area, scientists must act as all other citizens to encourage responsible action.

10. Conclusion to date

I have greatly enjoyed the hundreds of conversations that I have gotten to have with hydrologists and decision makers during my 2016 Darcy Lecture series. I think that the Darcy Lecture is a rare opportunity to have a conversation around the world and throughout a discipline. What's more, I think that these opportunities are increasingly important as science is called upon to help inform ever more societally important questions. It is not a task for everyone — there are costs to family and career for taking a year out of your life. But, if you are interested in expanding your thinking, promoting broader conversation, and simply seeing the world, there is no better experience that the Darcy Lecture or its cousins in other fields. This opportunity would not have been possible for me without the generous support of the many hosts who have made my visits possible. Most of all, my year has been possible due to the selfless support of my wife (Leslie) and my children (Ben and Luke) and the very generous contributions of time and good will by my fellow faculty members in the Department of Hydrology and Atmospheric Sciences at the University of Arizona.

11. Audience questions — Japan

I have maintained a blog throughout my tour. This provides access to supporting material, photos of general interest, and a record of all of the questions that I was asked (and my answers) throughout the year. I learned a great deal from these questions and discussions that followed from them. I would encourage you to visit the site (https://darcylecture2016.wordpress.com/ — the page 'Topics for Discussion' includes the questions) for more information. But, to give a flavor of the types of questions

that I was asked, I am including the questions that I was asked during my three stops in Japan. My answers can be found in this article and on the blog!

10/28/16 — audience member — Kyoto, Japan
— How do we form a model ensemble in a way that avoids bias?

10/28/16 — Masaru Mizoguchi — Tokyo University, Japan
— Following on your element of storytelling — I have spent a lot of time working on the problems related to Fukushima these five years. In this case, there are many facts and many people telling different stories. The problem is — even if we make a good, scientifically informed decision, many non-scientific people don't believe the scientific basis of the decision. Maybe this is most serious problem in science and technology caused by the Fukushima Daiichi nuclear disaster in Japan. So, I think that it is also important that we find better ways to communicate our science to the public to support decisions that have been made using science.

10/28/16 — Audience Member — Kyoto, Japan
— How did you choose the locations for the nine observations in your example?

10/26/16 — Ken Kawamoto — Saitama University, Japan
— Sometimes it can be unclear who the actual decision maker is. What do we do in these cases?

10/26/16 — Yosuke Matsuda — Mie University, Japan
— How does your approach relate to global climate change discussions?

10/24/16 — audience member — Tokyo, Japan
— How do you make a decision if you have multiple models that make different predictions? In particular, what do you do if you have multiple different groups who are involved in the decision making process?

10/24/16 — Satoshi Izumoto — Tokyo University, Japan
— When can you use statistical models and when do you need a physical model?

10/24/16 — Toshiko Komatsu — Saitama University, Japan
— It is not clear to me how your results for the contamination example relate to a specific decision that the client had to make. Can you clarify?

10/21/16 — Mitsuyoshi Ikeda — Nagasaki, Japan — Japanese Association of Hydrogeologists
— I appreciate your use of poetry in a scientific talk. I will return this with a question inspired by Shakespeare. I have over 30 years of experience in geologic modeling. I have found that the relationship between cost and improved understanding is not linear — there are times of very steep learning for relatively little cost and other times of little learning at high cost. So, how can we find the informative conditions? To measure or not to measure, that is the question!

References

Bredehoeft, J. (2005): The Conceptualization Model Problem — Surprise. Hydrogeology Journal, 13: 37–46.

Eagleman, D. (2012): Incognito: The Secret Lives of the Brain. First Vintage Books edition. Vintage Books, New York.

Goodwin, D.K. (2012): Team of Rivals: The Political Genius of Abraham Lincoln. Simon & Schuster, New York.

Gould, S.J. (1980): The Panda's Thumb: More Reflections in Natural History. Norton, New York.

Kahneman, D. (2011): Thinking, Fast and Slow. Farrar, Straus and Giroux, New York.

Kikuchi, C.P., Ferré, T.P.A. and Vrugt, J.A. (2015): On the optimal design of experiments for conceptual and predictive discrimination of hydrologic system models, Water Resour. Res., 51: 4454–4481, DOI:10.1002/2014WR016795.

Thaler, R.H. and Sunstein, C.R. (2009): Nudge: Improving Decisions About Health, Wealth, and Happiness. Penguin Books, New York.