Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology

James W. Kirchner[1]

Received 15 June 2005; revised 12 December 2005; accepted 16 December 2005; published 18 March 2006.

[1] The science of hydrology is on the threshold of major advances, driven by new hydrologic measurements, new methods for analyzing hydrologic data, and new approaches to modeling hydrologic systems. Here I suggest several promising directions forward, including (1) designing new data networks, field observations, and field experiments, with explicit recognition of the spatial and temporal heterogeneity of hydrologic processes, (2) replacing linear, additive "black box" models with "gray box" approaches that better capture the nonlinear and non-additive character of hydrologic systems, (3) developing physically based governing equations for hydrologic behavior at the catchment or hillslope scale, recognizing that they may look different from the equations that describe the small-scale physics, (4) developing models that are minimally parameterized and therefore stand some chance of failing the tests that they are subjected to, and (5) developing ways to test models more comprehensively and incisively. I argue that scientific progress will mostly be achieved through the collision of theory and data, rather than through increasingly elaborate and parameter-rich models that may succeed as mathematical marionettes, dancing to match the calibration data even if their underlying premises are unrealistic. Thus advancing the science of hydrology will require not only developing theories that get the right answers but also testing whether they get the right answers for the right reasons.

Citation: Kirchner, J. W. (2006), Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology, *Water Resour. Res.*, *42*, W03S04, doi:10.1029/2005WR004362.

[2] The day-to-day business of hydrology has largely been shaped by the need to solve practical problems, such as predicting floods and droughts, managing water resources, and designing water supply infrastructure. As a result, many hydrologists view their objective as developing practical predictive tools for operational purposes. For many routine operational purposes one just needs methods that get the right answers; for example, one might just need methods that predict stream flows, groundwater levels, or water quality with sufficient accuracy for the task at hand. From that operational perspective many contemporary approaches may be good enough for many practical purposes. However, to advance the science of hydrology, as opposed to the operational practice of hydrology (that is, to improve our understanding of how hydrologic systems work), we need to know whether we are getting the right answers for the right reasons. Furthermore, getting the right answers for the right reasons could be crucial for getting the right answers at all, if conditions shift beyond our range of prior experience (due to extreme precipitation events, climate change, or shifts in land use, for example).

[3] My focus in this commentary will be on advancing hydrologic science, rather than providing better predictions for operational purposes, although of course one hopes that the former may lead to the latter. My objective is to ask how we can develop better hydrologic measurements, analyses, and models, in order to more consistently get the right answers for the right reasons. I am neither the first commentator on this subject, nor the most luminous [e.g., *Klemes*, 1986, 1988; *Grayson et al.*, 1992; *Beven*, 2002]. Therefore I will make no particular claim of originality for the remarks presented here, but can only hope that they are framed in useful ways.

[4] In my view, advancing the science of hydrology will require new hydrologic measurements, new methods for analyzing hydrologic data, and new approaches to modeling hydrologic systems. These three essential aspects of hydrology will all be advanced if we take full advantage of the linkages between them. Some promising directions forward, in my view, include (1) designing new data networks, field observations, and field experiments, explicitly recognizing the spatiotemporal heterogeneity of hydrologic processes, (2) developing "gray box" data analysis methods that are more compatible with the nonlinear, nonadditive character of hydrologic systems, (3) developing physically based governing equations for hydrologic behavior at the catchment or hillslope scale, recognizing that they may look different from the equations that describe the small-scale physics, (4) developing models that are minimally parameterized, and therefore stand some chance of failing the tests that they are subjected to, (5) developing ways to test models more comprehensively and incisively, given the

[1]Department of Earth and Planetary Science, University of California, Berkeley, California, USA.

0043-1397/06/2005WR004362$09.00

intrinsic limitations of the available data. I will expand on each of these in the comments that follow.

[5] Let me begin, as all science begins, with the subject of data. From the growing volume and sophistication of hydrological theorizing over the past several decades, one might lose sight of the fact that the ultimate source of hydrological information is field observations and measurements. It is worth considering the extent to which current hydrological understanding is constrained by the kinds of measurements that have heretofore been available, and how those constraints can be loosened by new measurement technologies and new strategies for their deployment.

[6] Our current hydrological measurement networks have been, with few exceptions, designed for operational purposes rather than scientific ones. In California's Sierra Nevada mountains, for example, stream gauges are typically located downstream of dams, making them nearly useless for understanding catchment processes. As another example, most rainfall observations are made where people live; thus steep terrain is inherently underrepresented, so we have limited observations documenting how topography shapes precipitation patterns. Then, of course, there is the well-known problem that the surface area of a standard rainfall collector is typically 8 or 10 orders of magnitude smaller than the catchment that it is intended to represent. The spatial heterogeneity of rainfall implies that conventional precipitation measurements cannot be spatially representative; as a result, time series data from many small catchments contain occasional high-flow events that appear to have occurred in the absence of rainfall, simply because convective storm cells have missed the rainfall collector. Next to precipitation, the dominant term in the water balance is often evapotranspiration rather than runoff, but measurements of evapotranspiration rates are very scarce indeed. One could go on with examples like these.

[7] In the future, these kinds of data constraints should be alleviated somewhat by new measurement technologies, and by new hydrologic observatory networks. However, these new investments in hydrologic measurement infrastructure are likely to be expensive, and hydrologists will need to make a convincing case for them. Doing so will entail documenting what, specifically, hydrologists will be able to do with these new observations, that they cannot do without them. A useful template for making such a case is the report by the *Committee on Earth Gravity from Space* [1997]. That report documented, in quantitative terms, how satellite observations of subtle variations in Earth's gravitational field could be used to measure a broad range of geophysical phenomena, including seasonal changes in soil moisture and groundwater storage. It also quantified the tradeoffs between the design parameters of such a satellite system. The report was noteworthy in that it used simple models to extrapolate from what was known about the measurement technologies and the phenomena to be measured, to quantitatively demonstrate what could be achieved with a new measurement system. It was also noteworthy in its relative lack of the usual platitudes about "revolutionary new science" or "groundbreaking advances for the 21st Century" or the like. Furthermore, and most importantly, the report was unequivocally successful; it launched the GRACE (Gravity Recovery and Climate Experiment) satellite mission, with international support totaling over 100 million dollars. One can hope that in the near future, hydrologists will be able to make a similarly convincing case for major investments in new measurement infrastructure. However, for the moment, many hydrologists recognize that most of the existing hydrologic data are less than ideal for scientific purposes.

[8] What is less widely recognized, however, is that these data are typically analyzed with mathematical tools that may be inherently ill suited to hydrologic systems. For example, students studying hydrology often learn a suite of mathematically elegant methods (unit hydrographs, autoregressive models, and so forth) that assume that hydrologic systems are linear and additive. However, typical hydrologic systems are typically nonlinear and nonadditive, often dramatically so.

[9] Stuffing nonlinear, nonadditive systems into linear, additive black boxes can produce awkward results, no matter how artfully the mathematics is done. Just as any curve will be nearly linear over small enough segments, these approaches can sometimes give good enough answers as long as the system does not stray too far from the range of conditions represented by the calibration data. However, when the system is driven far beyond "normal" conditions (which is precisely when the answers matter most) these approaches often become unreliable. They don't extrapolate well, because their underlying premises were selected for mathematical convenience rather than physical realism. We need data analysis methods that are better informed by hydrological insight [e.g., *Wittenberg*, 1999; *Young*, 2003], to supersede the linear black box methods that we know are inconsistent with the fundamental mechanisms underlying many hydrologic processes. Getting the right answers for the right reasons will require "gray box" data analysis tools, ones that contain enough hydrological realism to capture the nonlinear, nonadditive behavior of hydrological systems.

[10] Getting the right answers for the right reasons has also been a rallying cry for those who would reject empirical approaches entirely, in favor of physically based mechanistic models of hydrologic systems. Surely, so the argument goes, if we model hydrologic systems using physically based governing equations (e.g., Darcy's law, Richards' equation, or the advection-dispersion equation), and if we get the right answers, then we must be getting the right answers for the right reasons. Yet when such models are calibrated to data from one time interval, they often perform poorly when tested against data from another time interval with different patterns of rainfall forcing [e.g., *Seibert*, 2003]. Similarly, models that are developed for one catchment often perform poorly when tested against data from other catchments. This casts doubt on the models' ability to predict how catchments will respond as conditions change.

[11] It is almost axiomatic that we need "physically based" models in order to make reliable predictions beyond the range of prior observations. However, the key question is not whether models of hydrologic systems should be physically based; instead, the question is how they should be based on physics. The physical laws governing water movement at small scales have been understood for decades. What we still don't understand well enough is how to apply these physical laws to systems that are complex,

heterogeneous on all scales, and often poorly characterized by direct measurement.

[12] To date, most "physically based" models of hydrologic systems are based on an implicit upscaling premise. This premise assumes that the microphysics in the heterogeneous subsurface will "scale up" such that the behavior at the model's grid scale will be described by the same governing equations (e.g., Darcy's law, Richards' equation), with state variables (e.g., water flux, volumetric water content, hydraulic potential) that are averaged, and with "effective" parameters (e.g., saturated conductivity, characteristic curves) that somehow subsume the heterogeneity of the subsurface. By necessity, these effective parameters are estimated through model calibration.

[13] Such models are often good mathematical marionettes; they often can dance to the tune of the calibration data. However, their predictive validity is often in doubt [*Seibert*, 2003]. This is because these models are usually overparameterized, in the sense that many different parameter sets will give almost identical fits to the calibration data (the equifinality problem [*Beven and Binley*, 1992]), but can yield dramatically different predictions of how the system will behave as conditions change. This is not an algorithmic problem, and it does not have an algorithmic solution. Instead, this problem arises because the free parameters in typical catchment models create many (many, many) more dimensions of flexibility than typical calibration data sets can constrain. Model parameters for an individual catchment will often vary significantly depending on the particular time interval that is used for calibration, and many models cannot accurately represent low-flow and high-flow behavior with a single set of parameter values [*Wagener*, 2003]. This violates a basic principle of mathematical modeling, namely, that the constants should stay constant while the variables vary.

[14] This implies that there is something structurally wrong with the models. In my view, we should more seriously consider the possibility that the implicit upscaling premise outlined above may be wrong. That is, we should more carefully explore the hypothesis that the effective governing equations for such heterogeneous systems at large scale may be different in form, not just different in the parameters, from the equations that describe the small-scale physics.

[15] As an example, recent field observations by Jeff McDonnell and his collaborators suggest that in humid catchments, rainfall inputs create a patchwork of mobile water as individual points on hillslopes reach sufficient saturation to become highly conductive; these observations also suggest that storm response is controlled by how this patchwork merges together into a spatially connected network as the hillslope wets up (J. J. McDonnell, personal communication, 2005). It seems unlikely that this hydrologic response could be mimicked by applying the conventional governing equations to the hillslope as a whole, even with considerable flexibility in picking the values of the "effective" parameters. As another example, theoretical calculations show that downslope advection and dispersion should produce a distinctive traveltime distribution for rainfall inputs distributed across a hillslope [*Kirchner et al.*, 2001]. These calculations agree with observations [*Kirchner et al.*, 2000], but are markedly different from what one would obtain by applying the advection-dispersion equation to any single "effective" flow path length. These examples imply that the effective governing equations (at the scale of hillslopes, catchments, or typical model grid cells) may look different from the point-scale equations that we all learned in school. As the examples above suggest, insight into these effective governing equations can come both from direct field measurements that span a range of scales, and from theoretical studies that integrate the point-scale physics across spatial domains.

[16] Whereas the problem of parameter identification has been emphasized in the hydrologic literature, the more fundamental (and difficult) problem of structural identification has received less attention than it deserves [*Butts et al.*, 2004]. Likewise, whereas many hydrologists recognize that overparameterization makes parameter identification problematic, it is less clearly understood that overparameterization also makes structural identification difficult. Parameter tuning makes models more flexible, and thus makes their behavior less dependent on their structure. This in turn makes validation exercises less effective for diagnosing models' structural problems. By making it easier for models to get the right answer, overparameterization makes it harder to tell whether they are getting the right answer for the right reason.

[17] It is not widely recognized that very little parameter tuning is still too much. For example, *Jakeman and Hornberger* [1993] showed that typical rainfall-runoff data only contain enough information to constrain simple hydrological models with up to four free parameters. Similarly, even with detailed hydrological and geochemical time series, *Hooper et al.* [1988] were unable to constrain a simple six-parameter model. By contrast, many catchment models have dozens of free parameters. Each additional parameter represents a whole new dimension of parameter space, so the overparameterization problem grows rapidly and nonlinearly with the number of free parameters.

[18] These considerations imply that in order to know whether we are getting the right answers for the right reasons, we will need to develop reduced-form models with very few free parameters. One promising avenue forward may be the approach outlined by *Kirchner et al.* [2001], which seeks a middle path between lumped-parameter and spatially distributed models. These "middle path" models attempts to capture the spatially extended character of hillslopes and catchments directly in their governing equations, without requiring explicit spatial disaggregation and the accompanying proliferation of free parameters.

[19] No matter what form the emerging theory will take, it is the collision of theory and data that forms the core of scientific advance in any field, including hydrology. The collision of theory and data will be more scientifically productive if we can develop models that are more parametrically efficient (and therefore less immune to being proven wrong), and if we can develop model testing regimens that compare models against data more incisively [*Kirchner et al.*, 1996]. Devising more incisive tests requires recognizing that we want models to be able to predict catchment responses to particular types of external forcing, such as climate change or land use change. We therefore need to test how well models can make those kinds of predictions, which is different from testing how nicely a

mathematical marionette can dance to a tune it has already heard.

[20] For example, typical split sample tests (in which a model is calibrated against one time period and tested against another) are not very revealing, because the two time series typically represent similar conditions and exhibit similar behavior. These are weak tests because the model is being tested against a data set that is functionally equivalent to the one that it has already been calibrated with. Instead, we should be making wider use of differential split sample tests, in which the calibration and validation data sets represent different conditions and different behaviors. These are more incisive tests, particularly if the calibration and validation data sets are different in ways that reflect the external forcings that are of particular concern, such as climate change. Differential split sample tests are rare, and models often fail them, but these failures are often highly instructive [e.g., *Seibert*, 2003].

[21] Models are often compared with data by simply overlaying the model predictions onto the observed time series data, or by plotting the predictions against the observed values. Such comparisons often fail to diagnose serious model deficiencies, either because the forcings of interest are obscured in the observational data, or because the model's fit to the data mostly reflects mechanisms other than the ones that control how it responds to the forcings of interest. Such cases call for a synthesis of modeling and data analysis. In this approach, one first statistically extracts the relationships of greatest relevance from the observations, correcting for other confounding factors. One then tests these relationships directly against the model; for an example, see *Kirchner et al.* [1996]. A variant of this approach uses statistical methods to extract the relationships of interest from both the observational data and the modeled time series (correcting for confounding factors in both), and then compares them against one another.

[22] Chemical and isotopic data provide another powerful test of whether we are getting the right answers for the right reasons. Many models can predict the rapid hydrograph response that is commonly observed in small catchments. However, small catchments also commonly exhibit strong damping in passive tracers (implying that storm runoff is predominantly ''old'' water), and exhibit strong concentration-discharge relationships in reactive tracers (implying that the ''old'' water that is mobilized during stormflow is chemically different from the ''old'' water that constitutes base flow). These three patterns of behavior provide important constraints on hypothesized mechanisms for stormflow generation [*Kirchner*, 2003]. Furthermore, because they characterize many different catchments, these patterns demand a general explanation: an elegant theory rather than an elaborate, highly parameterized megamodel.

[23] The full potential of chemical and isotopic data has not yet been achieved, because chemical measurements are typically made only weekly or monthly, whereas catchments often respond hydrologically and chemically on timescales of minutes to hours. Inferring the hydrochemical behavior of catchments from weekly or monthly samples is like trying to understand a symphony when one can hear just one note every minute or two. In the past, the costs of sampling and analysis made it difficult to collect high-frequency chemical and isotopic time series data. Recent technological advances are now loosening those constraints, and high-frequency sampling is revealing richly textured chemical behavior that had previously been hidden from view [*Kirchner et al.*, 2004]. As more detailed hydrochemical data sets become available, they will prove to be instrumental in developing the next generation of hydrologic theory.

[24] Although much of this commentary has focused on mathematical modeling and analysis of data, it is worthwhile to return to where this discussion began, and to emphasize again that all hydrological knowledge ultimately comes from observations, experiments, and measurements. These create all the information contained within hydrological data, and mathematical tools can at best only clarify (and at worst, obscure or distort) the information those data contain. Field observations, in particular, provide direct insights into processes and thus are crucial to the development of better hydrological theories. Manipulation experiments can provide particularly incisive tests of hydrological theory, because they can create experimental conditions that differ substantially from historical data, and because controlled experiments can isolate individual mechanisms, thus providing a more precisely defined 'target' for the theory to hit. Thus the advancement of hydrological modeling and analysis ultimately depends on supporting new experimental work, new field observations, and new data collection networks [*Grayson et al.*, 1992; *Hornberger and Boyer*, 1995].

[25] We need to develop better models and better analysis tools; we also need to create better data to model and analyze. Rethinking our theories and how we test them will bring major benefits in the long run. Striving not simply to get the right answers, but to get the right answers for the right reasons, will be both challenging and illuminating.

[26] **Acknowledgment.** I thank Xiahong Feng, Colin Neal, Beth Boyer, Christina Tague, and Sarah Godsey for many helpful discussions on these issues; Rick Hooper, Julia Jones, and Roger Bales for their reviews of the manuscript; and the National Science Foundation (EAR-0125550) for financial support.

References

Beven, K. (2002), Towards a coherent philosophy for modelling the environment, *Proc. R. Soc. London, Ser. A*, *458*, 2465–2484.

Beven, K. J., and A. M. Binley (1992), The future of distributed models: Model calibration and predictive uncertainty, *Hydrol. Processes*, *6*, 279–298.

Butts, M. B., J. T. Payne, M. Kristensen, and H. Madsen (2004), An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation, *J. Hydrol.*, *298*, 242–266.

Committee on Earth Gravity from Space (1997), *Satellite Gravity and the Geosphere*, 126 pp., Natl. Acad. Press, Washington, D. C.

Grayson, R. B., I. D. Moore, and T. A. McMahon (1992), Physically based hydrologic modeling: 2. Is the concept realistic?, *Water Resour. Res.*, *28*, 2659–2666.

Hooper, R. P., A. Stone, N. Christophersen, E. de Grosbois, and H. M. Seip (1988), Assessing the Birkenes model of stream acidification using a multisignal calibration methodology, *Water Resour. Res.*, *24*, 1308–1316.

Hornberger, G. M., and E. W. Boyer (1995), Recent advances in watershed modeling, *U.S. Natl. Rep. Int. Union Geod. Geophys. 1991–1994*, *Rev. Geophys.*, *33*, 949–957.

Jakeman, A. J., and G. M. Hornberger (1993), How much complexity is warranted in a rainfall-runoff model?, *Water Resour. Res.*, *29*, 2637–2649.

Kirchner, J. W. (2003), A double paradox in catchment hydrology and geochemistry, *Hydrol. Processes*, *17*, 871–874.

Kirchner, J. W., R. P. Hooper, C. Kendall, C. Neal, and G. Leavesley (1996), Testing and validating environmental models, *Sci. Total Environ.*, *183, 33–47.*

Kirchner, J. W., X. Feng, and C. Neal (2000), Fractal stream chemistry and its implications for contaminant transport in catchments, *Nature*, *403*, 524–527.

Kirchner, J. W., X. Feng, and C. Neal (2001), Catchment-scale advection and dispersion as a mechanism for fractal scaling in stream tracer concentrations, *J. Hydrol.*, *254*, 81–100.

Kirchner, J. W., X. H. Feng, C. Neal, and A. J. Robson (2004), The fine structure of water-quality dynamics: The (high-frequency) wave of the future, *Hydrol. Processes*, *18*, 1353–1359.

Klemes, V. (1986), Delettantism in hydrology: Transition or destiny?, *Water Resour. Res.*, *22*, S177–S188.

Klemes, V. (1988), A hydrological perspective, *J. Hydrol.*, *100*, 3–28.

Seibert, J. (2003), Reliability of model predictions outside calibration conditions, *Nord. Hydrol.*, *34*, 477–492.

Wagener, T. (2003), Evaluation of catchment models, *Hydrol. Processes*, *17*, 3375–3378.

Wittenberg, H. (1999), Baseflow recession and recharge as nonlinear storage processes, *Hydrol. Processes*, *13*, 715–726.

Young, P. (2003), Top-down and data-based mechanistic modelling of rainfall-flow dynamics at the catchment scale, *Hydrol. Processes*, *17*, 2195–2217.

J. W. Kirchner, Department of Earth and Planetary Science, 307 McCone Hall, 4767, University of California, Berkeley, CA 94720, USA. (kirchner@seismo.berkeley.edu)