Clark, Martyn P.; Kavetski, Dmitri; Fenicia, Fabrizio

Reply to comment by K. Beven et al. on "Pursuing the method of multiple working hypotheses for hydrological modelling"

Water Resources Research, 2012; 48(11):W11802

Reply to comment by K. Beven et al. on "Pursuing the method of multiple working hypotheses for hydrological modeling"

Martyn P. Clark,[1] Dmitri Kavetski,[2] and Fabrizio Fenicia[3]

Received 15 June 2012; revised 10 September 2012; accepted 6 October 2012; published 6 November 2012.

Citation: Clark, M. P., D. Kavetski, and F. Fenicia (2012), Reply to comment by K. Beven et al. on "Pursuing the method of multiple working hypotheses for hydrological modeling", *Water Resour. Res.*, *48*, W11802, doi:10.1029/2012WR012547.

1. Introduction and Scope

[1] We thank *Beven et al.* [2012], hereafter referred to as B12, for taking the time to comment on our opinion paper [*Clark et al.*, 2011a]. We are pleased that our paper piqued their interest, and we are pleased that B12 agree with much of what we say. We also welcome the opportunity to elaborate on our critique of the Generalized Likelihood Uncertainty Estimation (GLUE) methodology.

[2] The B12 comment sets the stage for some interesting discussion and debate. While the B12 comment is primarily focused on expressing their opinions about the superiority of GLUE over Bayesian approaches, the comment provides a good summary of many important challenges in hydrological sciences, including the B12 perspective on hypothesis-testing. In doing so, the B12 comment brings to the forefront the fundamental issues that we *must* address as we collectively seek to improve the fidelity of our models.

[3] Our response below highlights the need for a carefully controlled approach to model evaluation. This was one of the central aims of our opinion paper, where we present a methodology which entails both (1) isolating the constituent hypotheses in a model (e.g., experimenting with different options for specific processes and/or scaling behavior, while keeping all other components of the model fixed); and (2) using the available data and physical insights in creative ways to scrutinize different modeling alternatives. Elements of this methodology have been applied in several recent studies using a mix of qualitative and quantitative diagnostics [e.g., *Clark et al.*, 2011b; *Kavetski et al.*, 2011], and using the extended GLUE framework [e.g., *Krueger et al.*, 2010]. The controlled approach to model evaluation we advocate in our opinion paper—and in the response below—requires a combination of multiple tools and strategies to pursue the several distinct aspects of this methodology, with Bayesian methods being one of these tools.

[4] Our response identifies a great deal of common ground between our opinions and the sentiments expressed in B12. Although B12 may disagree with our choice of methods, including standard probability theory, statistics, and Bayesian techniques, we do share several broader aims and perspectives. In particular we agree with B12 that "We want a tool that will be useful in simulation or prediction and that reflects our qualitative perceptual knowledge of real-world processes." We are hence confident that our exchange adds to the ongoing constructive discussion on the suitability of different model analysis strategies, and helps define tractable ways forward for those interested in improving the process of model development and evaluation.

[5] Our response to the B12 comment is structured as follows. First, we review the B12 summary of the major science challenges in hydrological model analysis and consider how these challenges are pursued in our multiple hypothesis methodology. Second, we question the B12 defense of GLUE, pointing out that the aspects in which GLUE actually differs from Bayesian methods do not address the real challenges of an inference framework. Rather they simply weaken its descriptive, predictive and diagnostic capabilities, including the rigor with which model hypotheses can be tested. Third, we discuss the B12 perspectives on our modeling approach, emphasizing the importance of controlled approaches to model rejection and the need for subjectivity when stronger knowledge is not available. In responding to B12, our aim is to encourage hydrologists to think more critically of the fundamental premises, assumptions and limitations of different model analysis methodologies.

2. B12 Discussion of the Major Science Challenges

[6] We agree with B12 when they articulate many of the research challenges facing the community, including (1) the presence of epistemic errors due to limited process understanding/representation and data limitations; (2) the danger of spuriously rejecting model hypotheses simply because of data errors, and (3) avoiding overfitting when the various sources of error cannot be characterized or distinguished in a meaningful way. These are challenges for any modeling framework, whether based on theoretical arguments or on ad hoc considerations. The purpose of our opinion paper was to analyze tractable strategies that will allow the community to address some of these outstanding modeling challenges in a systematic and carefully controlled way.

[1] Research Applications Laboratory, National Center for Atmospheric Research, Boulder, Colorado, USA.

[2] School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, South Australia, Australia.

[3] Centre de Recherche Public—Gabriel Lippmann, Belvaux, Luxembourg.

Corresponding author: M. P. Clark, Research Applications Laboratory National Center for Atmospheric Research, Boulder, CO 80301, USA. (mclark@ucar.edu)

0043-1397/12/2012WR012547

2.1. Epistemic Errors

[7] We agree with B12 when they emphasize the need to study epistemic errors in more detail – indeed, the very objective of scientific hydrology is to investigate and reduce epistemic errors. This includes improving process understanding through a combination of fieldwork and modeling, avoiding oversimplified models, identifying inadequate data (e.g., a sparse rain gauge network where all stations record zero precipitation during some storm events), and other tasks. Furthermore, given inevitable approximations in environmental modeling, it is the very goal of uncertainty analysis to quantify these approximation errors as meaningfully as possible. For example, *Gupta et al.* [2012] argue that a major modeling challenge is in detecting epistemic errors, characterizing their impact, attributing their cause, and correcting them, suggesting "it is to the problem of underlying epistemic cause (things we could in principle, but do not in practice, know) that our collective brainpower, curiosity and investigation must be directed."

[8] The merits of different strategies to characterize lack of knowledge are the subject of ongoing debates in the broader science community, with discussions involving compelling qualitative and quantitative arguments, and with many distinct formalisms being proposed, including Bayesian methods, fuzzy theory, Dempster-Shafer theory, and others. All have their strengths and weaknesses, and the debates are unlikely to settle any time soon. The proposition by B12, elaborated in their earlier opinion pieces [e.g., *Beven*, 2006; *Beven et al.*, 2011; *Beven and Westerberg*, 2011], that probability theory and statistics are unsuitable for representing, or even approximating, lack of knowledge is just one of several views, rather than some established and accepted result.

[9] In our opinion, the presence of epistemic uncertainty does not require abandoning probabilistic frameworks. Ascertaining whether or not a particular source of uncertainty has been adequately characterized, or, more accurately, *approximated*, probabilistically is best pursued using quantitative hypothesis-testing (where different descriptions of the errors are hypothesized and tested a posteriori), rather than a priori by attempting to classify the error as epistemic or aleatory. As a simple example, calibration and validation provide direct quantitative estimates of the total errors (whether or not they are epistemic!), and modelers are free to experiment and test different probabilistic descriptions of these errors. These error descriptions can then be improved, including, where relevant, detecting and representing various nonstationarities, distinguishing between data of different quality, and, where necessary, censoring data of particularly poor quality. What B12 describe as "disinformative" data is also amenable to such analysis. As these improvements are carried out, and this may clearly take considerable effort, the potential for overestimating the information content of the data ("overconditioning") is reduced—again, whether or not the errors are epistemic or aleatory! Put simply, the more data we have, and the better its quality, the more we can begin characterizing and distinguishing different sources of error.

[10] A critical point here—overlooked in the B12 comment—is that the problem of epistemic errors is clearly much broader than a statistical one (see also the discussion in *Sivapalan* [2009]). While *characterizing* errors, including epistemic errors, is achieved through uncertainty analysis, *reducing* epistemic errors, including the effects of nonstationarites, is achieved through model improvements (to the extent that our process understanding and observational capabilities allow it). Our view is that investigating epistemic errors requires a more controlled—and more thoughtful—approach to model development and evaluation. In our opinion, it requires moving away from "blunt" approaches that do not identify specific model weaknesses, and instead adopting more incisive model analysis methods, in particular, the multiple hypothesis approach described in our opinion paper.

[11] The key message of our opinion paper is not Bayesian statistics—rather it is the need to decompose a model into its constituent hypotheses, including both the representation of individual processes and representations of how different processes combine to create the system-scale response—and attempting, inasmuch as possible, to evaluate each hypothesis separately, accounting for data uncertainty. In our opinion this provides the best opportunity to diagnose and remedy model deficiencies (epistemic error), without obscuring them by compensatory errors in different parts of the model. It also allows meaningfully attributing specific changes in the model behavior and predictive ability to specific, controlled changes in the model structure, without being confounded by a multitude of uncontrolled differences.

[12] Applying any model in prediction, especially in extrapolation, is certainly susceptible to system nonstationarity. This is evidently case specific. While in some cases existing nonstationarities could be detected from available data and reflected in the model structure and/or statistical error description, our only hope to anticipate genuine shifts in environmental behavior lies in process understanding, and in adequately estimating future forcing conditions.

2.2. Accounting for Data Errors and Avoiding Overfitting

[13] We agree with B12 that it is important to avoid "rejecting models that might be useful in prediction simply because of errors in the input data and evaluation observations." Indeed, we state in our opinion paper: "Practical issues such as data availability and data quality necessarily affect the insights that can be gained in a particular hydrological system—put simply; data uncertainty constrains our ability to discriminate among competing hydrological hypotheses."

[14] In our opinion, the *only* way to avoid this problem is to pursue a better quantification of data errors [e.g., see *Renard et al.*, 2011] and to collect new independent data whenever possible (including developing new measurement technologies). We have an entire section in our opinion paper [*Clark et al.*, 2011a] dedicated to stringent model diagnostics and the "clever use of data." We also emphasize statements made by others in the community that the use of streamflow data in model evaluation needs to go beyond simply computing sum of squared differences between simulated and observed streamflow (a point also stressed by Beven in the "limits of acceptability" papers). Moreover, we clearly acknowledge that the data and insights necessary for meaningful model evaluation are currently only available in a few research catchments. This represents a major challenge

for the community, which must be addressed through collecting additional data in a range of different hydrologic settings.

3. B12 Defense of GLUE

[15] The central claims of B12 are (1) Bayesian approaches to statistical inference are a special case of GLUE; and (2) Bayesian approaches make strong and unjustified assumptions about the error characteristics – although it is unclear to us what B12 view as the assumptions of GLUE, and whether (and why) they view these assumptions as justified. On the basis of these claims, B12 propose that GLUE is a more "common sense" approach to model evaluation and hypothesis testing, and claim that the extended GLUE methodology (termed the "limits of acceptability" approach), *unlike* Bayesian methods, already offers concrete avenues for resolving the "nonideal" modeling problems facing the hydrological modeling community. Here we consider the differences and similarities between GLUE and Bayesian approaches, and question whether GLUE is suitable for rigorous model inference, evaluation and predictive use.

3.1. Is GLUE Beginning to Move Toward Bayesian Methods?

[16] Consider the differences and similarities between Bayesian methods and GLUE. In Bayesian analysis, the posterior distribution of inferred quantities θ given observed data D and a prior distribution $p(\theta)$ is given by Bayes equation, $p(\theta|D) \propto p(D|\theta)p(\theta)$. In GLUE, the likelihood function $p(D|\theta)$ is replaced by a pseudo-likelihood function $f_{PL}(D,\theta)$, yielding $f_{GLUE}(\theta,D) \propto f_{PL}(D,\theta)p(\theta)$. The difference is much deeper than just what function to plug into the likelihood. While in the Bayesian methods the predictions are computed in a way that is consistent with the selected likelihood function (e.g., including residual errors, parameter stochasticity, etc.), virtually all GLUE studies have relied exclusively on parametric uncertainty (i.e., the ensemble of "behavioral" parameter sets) to compute the prediction limits. Another key difference is that while Bayesian methods allow the individual approximations and assumptions made when formulating the likelihood function to be scrutinized and improved a posteriori, we are not aware of how to disentangle, scrutinize and relax the various assumptions hidden within the pseudo-likelihood functions typically used in GLUE studies. In our opinion, it is irrational for B12 to claim that Bayesian methods are flawed because some particular applications have used "unrealistic" likelihood functions, given that a key component of Bayesian analysis is posterior diagnostics to evaluate and improve the adequacy of the likelihood function. The next sections elaborate on these points.

3.1.1. Bayesian Approach

[17] A Bayesian modeler would not claim to have the "correct" likelihood in any particular application, much like it would take a brave hydrologist to claim they have the "correct" hydrological model. Instead, the explicit aim of "formal" Bayesian modeling is to try to get as close as possible to this ideal – again, just like a hydrologist would like their model to be as realistic as possible. This is accomplished by iterating the following steps: (1) explicitly formulate all the knowledge and assumptions that the modeler is prepared to use when formulating the likelihood function and the prior. This may include data error analysis [e.g., *Renard et al.*, 2011], structural error representations [e.g., *Bulygina and Gupta*, 2011], and is not limited to additive error models [e.g., *Renard et al.*, 2011]. Mixtures and combinations of time- and/or space-varying (i.e., nonstationary) distributions can also be used, e.g., if residual error analysis or any other considerations suggest that the properties of errors change in time and/or in space; (2) investigate the posterior distribution, e.g., using sampling; (3) apply posterior diagnostics, including predictive tests using independent data, to scrutinize these assumptions; and (4) where appropriate, revise the likelihood function, collect new data, and, in case of incompatibilities, question the prior knowledge. The iterated application of steps (1)–(4) follows the basic principles of scientific hypothesis-testing, where modelers propose hypotheses (assumptions), and test them against available evidence.

[18] All these steps have been thoroughly documented in widely available Bayesian references, such as *Box and Tiao* [1973]. The fact that these steps are often poorly followed in conceptual hydrological modeling gives B12 every right to question these hydrological applications and their conclusions. But before their critique of the Bayesian principles can be taken seriously, it should directly challenge steps (1)–(4) and provide a convincing explanation of why they see it beneficial to short circuit these steps by inventing concepts such as "pseudo-likelihood" at the expense of foregoing the aim that prediction limits be quantitatively interpretable. We will return to these issues in section 3.2.1.

3.1.2. GLUE Approach

[19] In contrast to Bayesian modelers, B12 appear to presume that epistemic uncertainty is impossible to characterize using probability theory. Some of their concerns appear misplaced. For example, probability theory and statistics are not limited to the uncorrelated Gaussian distribution (a point we return to in section 4.2). B12 also appear to question key identities of probability theory, such as the multiplicative conditional probability equation $p(AB) = p(A|B)p(B)$, from which Bayes equation and other key relations are derived [*Ang and Tang*, 2007]. However, apart from its rather categorical nature, this B12 conjecture does not explain, let alone prove, why the so-called "informal" likelihood functions recommended in GLUE publications, such as the Nash-Sutcliffe index, are any better suited to this task. What, if any, are the requirements imposed by the GLUE methodology on the "pseudo"-likelihood function, how are these requirements different from the requirements in Bayesian likelihoods, and how does the GLUE methodology test and better satisfy these requirements?

[20] More importantly from a hypothesis-testing perspective, what assumptions should a GLUE modeler test after they have specified their pseudo-likelihood function? For example, most GLUE studies (including most of the B12 publications prior to 2006) use the Nash-Sutcliffe index as the pseudo-likelihood. We would welcome a clear quantitative explanation from B12 of whether they believe the Nash-Sutcliffe index makes fewer assumptions than the uncorrelated constant-variance Gaussian error model used in the most primitive Bayesian schemes. We would also welcome a clear statement of what these assumptions are, how a GLUE modeler would go about testing them quantitatively, and, most importantly, how they could be systematically relaxed if the modeler is dissatisfied. In the absence

of such explanations, we reiterate that the assumptions underlying the GLUE inference and predictions are difficult to disentangle and are hence hidden from scrutiny. Although using the Nash-Sutcliffe *as if* it were a likelihood function could produce "dotty" plots that appear "reasonable" to a given modeler, this does not address the question of quantitative and iterative hypothesis testing.

3.1.3. "Extended" GLUE Approach

[21] Now let us elaborate on our opinion that GLUE is shifting toward mainstream Bayesian methods—a point where B12 clearly disagree. If we inspect the actual equations used in the more recent "extended" GLUE methodology (the limits of acceptability approach) we can see the gradual replacement of the Nash-Sutcliffe index with increasingly elaborate functions, including "rectangular" error measures [*Winsemius et al.*, 2009], "triangular" error measures [e.g., *Liu et al.*, 2009; *Westerberg et al.*, 2011] and "trapezoidal" error measures [*Blazkova and Beven*, 2009; *Krueger et al.*, 2010]. These measures are beginning to explicitly and quantitatively reflect particular insights and assumptions made by the modeler regarding the magnitude and distribution of data and model errors. Evidently such error measures are much closer to describing $p(D|\theta)$ than the rather arbitrarily picked Nash-Sutcliffe metric. We can begin seeing these error measures as error models, except that instead of Gaussian assumptions the "new" GLUE modelers are making uniform, triangular and trapezoidal assumptions. Just as in the simplest Bayesian schemes, such assumptions are still applied to the residual errors (because that's what the GLUE error measures are applied to), and they are still applied independently from time step to time step, hence ignoring the error autocorrelation. Finally, since the parameter distributions (weights) are kept fixed after calibration and the residual errors are neglected (we disagree with B12 that "they are usually considered implicitly"—what does this really mean, and how can this "implicit" treatment be scrutinized?), we question whether GLUE can represent any kind of nonstationarity at all.

3.1.4. Is GLUE a Simplified and Incomplete Application of Bayesian Principles?

[22] A remaining question is how to test the assumptions on data and model errors in the "new" GLUE methodology. Following a careful reading of the studies cited by B12, we did not see any posterior diagnostics applied to test the new pseudo-likelihoods. For example, how does a GLUE modeler check the "nonmultiplicative" GLUE methods for combining probabilities? Can the data be used to a posteriori suggest or refine the form of the pseudo-likelihood function (e.g., triangular)? What guidance, beyond standard statistical tests (which, according to *Beven* [2006], GLUE may not even be intended to satisfy!), is available to improve on these strong and, as yet, unjustified assumptions? The description of the pseudo-likelihood function as part of the "audit trail" mentioned by B12 is indeed essential documentation, but does not, in itself, provide any scrutiny of its assumptions. It is for this reason we stated in the opinion paper that new GLUE studies, while moving closer to Bayesian principles, are still some way from applying adequate posterior scrutiny to their assumptions.

[23] The absence of any synthetic testing is a particular concern. Have B12 investigated the properties of their limits of acceptability approach using synthetic data? Every published study confirms that unless the likelihood function provides an adequate description of the data and model errors, poor inference and prediction are obtained *even* for synthetic data [e.g., *Stedinger et al.*, 2008]. For example, in all the experiments carried out by *Montanari* [2005], GLUE underestimated the total uncertainty of the model predictions, in many cases by a substantial amount. Given the poor performance of the GLUE approach on "ideal" test cases, it is rather difficult to accept the B12 claim that their methods are a "more common sense approach" for "nonideal" modeling problems. In our opinion, published critiques such as *Mantovan and Todini* [2006] and *Stedinger et al.* [2008] have correctly demonstrated the fundamental theoretical and practical flaws of GLUE, and, while we appreciate some of their general motivation, we were not convinced by the arguments proposed in subsequent responses including *Beven et al.* [2008], *Beven and Westerberg* [2011] and others.

[24] It is this analysis of GLUE—old and new—that gives us reason to view GLUE as a simplified, incomplete application of some Bayesian principles. In our view, GLUE is an intermediate, largely ad hoc step between calibration schemes where the modeler simply specifies the objective function they want to use, and more rigorous probabilistic approaches that aim to formulate the objective function in a way that comes as close as possible to providing not just point estimates of the quantities of interest (parameters, predictions, etc.), but also quantitatively interpretable uncertainty estimates.

3.2. A "Common Sense" Approach to Model Evaluation?

[25] B12 make the claim that GLUE "does allow a more common sense approach to model evaluation and hypothesis testing." This claim deserves further scrutiny. In the following sections, we assess if the claim is supported by previously published papers on the limits of acceptability approach.

3.2.1. Is GLUE Suitable for Uncertainty Quantification?

[26] Consider the key purpose of quantitative uncertainty analysis methods—that is, to provide quantitative estimates of the uncertainty in the model predictions and/or quantitative estimates of the uncertainty in the model parameters. Yet the meaning and utility of the GLUE dotty plots as estimated parameter distributions is dubious, and the GLUE Manifesto acknowledges that the prediction limits in GLUE are not even *intended* to satisfy even basic requirements such as encompassing the relevant fractions of observations [*Beven*, 2006].

[27] If GLUE uncertainty estimates have no quantitative significance, then what is the point of affixing numerical values to these quantities? Can these quantities be somehow verified? What are their detailed underlying assumptions? And, more generally, what is the output of the GLUE methodology that is *at least intended* to be quantitatively interpretable, and what is its interpretation? Based on these considerations, is it really appropriate to view GLUE as a quantitative uncertainty analysis methodology? Here, we concur with *Montanari* [2007] that GLUE as such is a basic sensitivity analysis method, not an uncertainty analysis framework. The suggestion by B12 that Bayesian

methods are a special case of GLUE is a rather strong and perplexing misrepresentation.

3.2.2. Does GLUE in Itself Diagnose Model Weaknesses?

[28] In describing the "limits of acceptability" approach, B12 note that "posterior analysis of the residuals … may be a guide for model improvements or querying the usefulness of particular periods of calibration data." We agree, and indeed residual error analysis and data quality checks have been the staple of science and statistics for many decades. Their usefulness is precisely in helping detect outliers, questionable data, periods of poor model performance, etc.—thus directly testing the hypothesized hydrological model and any hypothesized data and structural error models.

[29] This leads to a more general question: does GLUE in itself provide any new tools for understanding model weaknesses, besides standard analysis of model residuals and data quality checks? Or is it the case that many of B12's hydrological insights can also be employed in more general approaches?

[30] We contend that effectively diagnosing model weaknesses requires controlling for compensatory errors in different parts of the model—a point made by many authors including *Kuczera and Franks* [2002], *Beven* [2006], and *Gupta et al.* [2008]. Specifically, we argue that the tools of statistical inference and prediction, including tests such as residual error analysis, will be most powerful when applied within a multiple hypothesis methodology that evaluates and compares alternative modeling decisions using multivariate data, fieldwork and process-oriented insights. At least one of the recent GLUE applications embarks on this path [e.g., *Krueger et al.*, 2010], and the key insights and advances presented there, in particular the testing of multiple modeling alternatives, are not GLUE-specific.

[31] Our main point here is that GLUE, in itself, does not protect against compensatory behavior. Rather, compensatory errors are problematic in any study reliant on aggregated measures of model performance, whether using "formal" or "informal" statistics. Indeed, the hydrological insights available in the more recent "limits of acceptability" papers [*Blazkova and Beven*, 2009; *Westerberg et al.*, 2011] cited in B12 as examples of hypothesis-testing using GLUE are also limited by compensatory errors. As an illustration, *Westerberg et al.* [2011] pose the question: "Are these … models then acceptable hypotheses of the hydrological processes in the respective catchments, or should they be rejected?", yet limit their comparison to differences between observed and simulated discharge, without inspecting internal model dynamics, other data, or process-oriented diagnostics that might better test the assemblage of hypotheses in their models. As such, we see the use of multiple data sources and multiple diagnostics [e.g., *Kuczera and Mroczkovski*, 1998; *Freer et al.*, 2004] as possible, indeed, highly desirable, in any modeling framework.

[32] Can we, as a community, do better hydrological hypothesis-testing? Yes, we *should*, and to do so requires more careful attention to all steps of the modeling process, including a more clear articulation of multiple working hypotheses and subhypotheses, more robust numerical implementation, more robust inference, and more incisive model diagnostics.

4. B12 Perspectives on Our Modeling Approach

[33] B12 have unfortunately misread our opinion paper in two important ways—they suggest that we are against model "rejection," and that we deny the role of "subjectivity." This reflects the narrow focus of B12 in terms of comparisons between GLUE and Bayesian methods, whereas our opinion paper presented a much broader and (in our view) a more balanced view of hypothesis testing.

4.1. Is Model Rejection a Good Thing?

[34] B12 state that model hypotheses will never actually be rejected in a formal Bayesian framework, because Bayes ratios only rank models relative to each other. And later in the paper they state that "…rejection, properly justified, is a good thing. It forces a reconsideration of what is causing the failure, which might be either the model hypotheses or disinformative data. It is then not really clear to us why the authors argue against a rejectionist framework, especially when they cite Popper in support of more rigorous hypothesis testing."

[35] The key point here is "properly justified." We are not arguing against model rejection per se – rather we are arguing against the blunt, "lumped" approach to model rejection that remains commonplace in the hydrological community. We seek to scrutinize individual hypotheses (to the extent that the data allows it), and select specific model components that best describe the processes they are intended to simulate. We point out that this approach "…is consistent with the philosophy employed … in TOPMODEL, which is presented as a set of concepts, rather than a fixed structure, based on the hypothesis that topography controls saturated areas and base flow." This approach avoids rejecting perfectly good modeling concepts. Later in the opinion paper we state "…by comparing model representations at the level of model subcomponents it becomes possible to select the best component hypotheses from different models, thereby avoiding the need to reject entire models (this makes better use of insights gained during model development)." We are therefore in favor of model rejection, but only when applied in a very precise way.

4.2. Subjectivity Perfectly Acceptable in the Absence of Stronger Knowledge

[36] We agree with B12 that it is necessary—and beneficial!—to incorporate subjective judgment into the model evaluation process. In fact, our opinion paper made many statements to this effect, including that "the apparent structural simplicity suggested in the absence of accurate quantitative ("hard") information should be judged against independent knowledge available from general hydrological theory and/or any qualitative ("soft") fieldwork evidence. In other words, model evaluation requires a mix of qualitative and quantitative insights."

[37] We do not claim that the Bayesian approach to hypothesis testing avoids subjectivity. On the contrary, the use of subjective knowledge is a key part of the Bayesian approach, and is actually what distinguishes Bayesian and frequentist methods. Unlike a frequentist, a Bayesian modeler is prepared to represent subjective knowledge using probability distributions. This has been the subject of immense debates in the statistical community, and is not necessarily universally accepted—many proponents of frequentist methods still disagree with using subjective knowledge in Bayes equation.

[38] Nor do we claim that merely formulating the likelihood function using a common probability distribution (e.g., the Gaussian distribution of residual errors) is "objective" or "formal." First, non-Gaussian distributions used in the extended GLUE methodology, such as uniform, triangular, and trapezoidal, etc., are neither "informal" nor "subjective"—they are just as "statistical" and suitable for Bayesian analysis as Gaussian distributions. Second, and much more importantly, being subjective in formulating hypotheses, if this means using qualitative insights, is not forbidden in science or in Bayesian statistics—rather the aim of "formal" methodologies is to explicitly formulate all assumptions and hypotheses, and apply focused and stringent diagnostics so that propositions that are not supported by the empirical evidence are improved or rejected. This rejection is conditional on the available data and prior knowledge and is hence not final—new evidence and/or data may require revisiting previously rejected hypotheses in a new light.

5. Concluding Remarks

[39] The early GLUE publications were among the first to recognize and widely critique the weaknesses of their contemporary calibration methods, and propose alternative approaches for investigating uncertainty in hydrological models [e.g., *Beven and Binley*, 1992; *Freer et al.*, 1996]. Their efforts have been followed up by generations of hydrologists and scientists in many other disciplines, and have helped to more seriously address the challenge of uncertainty in environmental modeling.

[40] As our community matures, it is important to continue improving model analysis strategies, taking advantage of ongoing advances in environmental physics, data collection and mathematical modeling. We enjoy active and productive collaborations with many in the GLUE community, and—though differences of opinion remain (as is acutely evident in this exchange)—these collaborations are resulting in tangible improvements in our model analysis methods. We agree with B12 when they state "…the issues raised here are not going to be resolved easily" and we agree that further developments and comparisons between different methods are of value. We hope that this comment/exchange helps define tractable ways forward for those interested in improving the process of model development and evaluation.

[41] **Acknowledgments.** We are grateful to John England, George Kuczera, Laurent Pfister, Ben Renard and David Rupp for their constructive criticisms of an earlier version of this manuscript.

References

Ang, A. H. S., and W. H. Tang (2007), *Probability Concepts in Engineering*, 2nd edition, Wiley, New York.

Beven, K. J. (2006), A manifesto for the equifinality thesis, *J. Hydrol.*, *320*, 18–36.

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

Beven, K. J., and I. K. Westerberg (2011), On red herrings and real herrings: Disinformation and information in hydrological inference, *Hydrol. Processes*, *25*, 1676–1680, doi:10.1002/hyp.7963.

Beven, K., P. Smith, and J. Freer (2008), So just why would a modeller choose to be incoherent?, *J. Hydrol.*, *354*, 15–32, doi:10.1016/j.jhydrol.2008.02.007.

Beven, K., P. J. Smith, and A. Wood (2011), On the colour and spin of epistemic error (and what we might do about it), *Hydrol. Earth Syst. Sci.*, *15*, 3123–3133.

Beven, K. J., P. J. Smith, I. K. Westerberg, and J. Freer (2012), Comment on Clark et al., Pursuing the method of multiple working hypotheses for hydrological modeling, W09301, 2011, *Water Resour. Res.*, doi:10.1029/2012WR012282, in press.

Blazkova, S., and K. Beven (2009), A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic, *Water Resour. Res.*, *45*, W00B16, doi:10.1029/2007WR006726.

Box, G. E. P., and G. C. Tiao (1973), *Bayesian Inference in Statistical Analysis*, John Wiley, New York.

Bulygina, N., and H. V. Gupta (2011), Correcting the mathematical structure of a hydrological model via Bayesian data assimilation, *Water Resour. Res.*, *47*, W05514, doi:10.1029/2010WR009614.

Clark, M. P., D. Kavetski, and F. Fenicia (2011a), Pursuing the method of multiple working hypotheses for hydrological modeling, *Water Resour. Res.*, *47*, W09301, doi:10.1029/2010WR009827.

Clark, M. P., H. K. McMillan, D. B. G. Collins, D. Kavetski, and R. A. Woods (2011b), Hydrological field data from a modeller's perspective. Part 2: Process-based evaluation of model hypotheses, *Hydrol. Processes*, *25*, 523–543, doi:10.1001/hyp.7902.

Freer, J., K. Beven, and B. Ambroise (1996), Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach, *Water Resour. Res.*, *32*, 2161–2173.

Freer, J. E., H. McMillan, J. J. McDonnell, and K. J. Beven (2004), Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures, *J. Hydrol.*, *291*(3–4), 254–277.

Gupta, H. V., T. Wagener, and Y. Liu (2008), Reconciling theory with observations: Elements of a diagnostic approach to model evaluation, *Hydrol. Processes*, *22*, 3802–3813.

Gupta, H. V., M. P. Clark, J. A. Vrugt, G. Abramowitz, and M. Ye (2012), Towards a comprehensive assessment of model structural adequacy, *Water Resour. Res.*, *48*, W08301, doi:10.1029/2011WR011044.

Kavetski, D., F. Fenicia, and M. P. Clark (2011), Impact of temporal data resolution on parameter inference and model identification in conceptual hydrological modeling: Insights from an experimental catchment, *Water Resour. Res.*, *47*, W05501, doi:10.1029/2010WR009525.

Krueger, T., J. Freer, J. N. Quinton, C. J. A. Macleod, G. S. Bilotta, R. E. Brazier, P. Butler, and P. M. Haygarth (2010), Ensemble evaluation of hydrological model hypotheses, *Water Resour. Res.*, *46*, W07516, doi:10.1029/2009WR007845.

Kuczera, G., and S. Franks (2002), Testing hydrologic models: Fortification or falsification?, in *Mathematical Modelling of Large Watershed Hydrology*, edited by, V. P. Singh and D. K. Frevert, Water Resour. Publ., Littleton, Colo.

Kuczera, G., and M. Mroczkovski (1998), Assessment of hydrologic parameter uncertainty and the worth of multiresponse data, *Water Resour. Res.*, *34*, 1481–1489, doi:10.1029/98WR00496.

Liu, Y., J. Freer, K. Beven, and P. Matgen (2009), Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error, *J. Hydrol.*, *367*, 93–103, doi:10.1016/j.jhydrol.2009.01.016.

Mantovan, P., and E. Todini (2006), Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology, *J. Hydrol.*, *30*, 368–381, doi:10.1016/j.jhydrol.2006.04.046.

Montanari, A. (2005), Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations, *Water Resour. Res.*, *41*, W08406, doi:10.1029/2004WR003826.

Montanari, A. (2007), What do we mean by uncertainty? The need for a consistent wording about uncertainty assessment in hydrology, *Hydrol. Processes*, *21*, 841–845.

Renard, B., D. Kavetski, E. Leblois, M. Thyer, G. Kuczera, and S. W. Franks (2011), Towards a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation, *Water Resour. Res.*, *47*, W11516, doi:10.1029/2011WR010643.

Sivapalan, M. (2009), The secret to 'doing better hydrological science': change the question!, *Hydrol. Processes*, *23*(9), 1391–1396, doi:10.1002/hyp.7242.

Stedinger, J. R., R. M. Vogel, S. U. Lee, and R. Batchelder (2008), Appraisal of the generalized likelihood uncertainty estimation (GLUE) method, *Water Resour. Res.*, *44*, W00B06, doi:10.1029/2008WR006822.

Westerberg, I. K., J.-L. Guerrero, P. M. Younger, K. J. Beven, J. Seibert, S. Halldin, J. E. Freer, and C. Y. Xu (2011), Calibration of hydrological models using flow-duration curves, *Hydrol. Earth Syst. Sci.*, *15*, 2205–2227, doi:10.5194/hess-15-2205-2011.

Winsemius, H., B. Schaefli, A. Montanari, and H. H. G. Savenije (2009), On the calibration of hydrological models in ungauged basins: A framework for integrating hard and soft hydrological information. *Water Resour. Res.*, *45*, W12422, doi:10.1029/2009WR007706.