Comment on "Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window" by Frank T.-C. Tsai and Xiaobao Li

Ming Ye,[1] Dan Lu,[1] Shlomo P. Neuman,[2] and Philip D. Meyer[3]

Received 7 August 2009; revised 2 December 2009; accepted 16 December 2009; published 11 February 2010.

Citation: Ye, M., D. Lu, S. P. Neuman, and P. D. Meyer (2010), Comment on "Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window" by Frank T.-C. Tsai and Xiaobao Li, *Water Resour. Res.*, *46*, W02801, doi:10.1029/2009WR008501.

[1] *Tsai and Li* [2008] assert that the Bayesian information criterion (BIC) [*Schwarz*, 1978] is better suited for comparing models having different parameters than is the Kashyap criterion (KIC) [*Kashyap*, 1982] because a Fisher information term in the latter may rank models with relatively large parameter estimation uncertainties higher than other models. We start by noting that KIC reduces asymptotically to BIC as the number of observations becomes large relative to the number of adjustable model parameters [*Ye et al.*, 2008]. If *Tsai and Li* [2008] were correct in their assertion, this would imply that it is better to treat a finite set of data as if it were theoretically infinite, a proposition that is logically unappealing and not necessary in practice.

[2] The Fisher information term imbues KIC with desirable model selection properties not shared by BIC [*Ye et al.*, 2008]: it sometimes prefers more complex models than does BIC because of its unique ability to discriminate between models not only on the basis of their goodness of fit to observational data and number of parameters but also on the quality of the available data and of the parameter estimates. To appreciate this role of the Fisher information term, it must not be considered in isolation as do *Tsai and Li* [2008] but rather in the context of all terms entering into KIC as do *Ye et al.* [2008]. The purpose of this comment is to elaborate on the discussion of *Ye et al.* [2008] by explaining further why the tendency of KIC to prefer models with relatively large parameter estimation uncertainty is a strength rather than a weakness.

[3] In a manner analogous to that of *Sivia and Skilling* [2006], we present a simple example which helps elucidate the role played by the Fisher information term in KIC and allows us to offer general observations regarding more complex applications, such as the groundwater inverse modeling analysis of *Tsai and Li* [2008]. Consider two models, A and B, having one adjustable parameter each, μ and λ, respectively. Bayes' theorem implies that the ratio between the posterior model probabilities, conditioned on an observation vector $\mathbf{D}$, is

$$\frac{p(A|\mathbf{D})}{p(B|\mathbf{D})}=\frac{p(\mathbf{D}|A)}{p(\mathbf{D}|B)}\frac{p(A)}{p(B)}, \tag{1}$$

where $p(A)$ and $p(B)$ are prior probabilities of the two models and

$$\begin{aligned} p(\mathbf{D}|A) &= \int p(\mathbf{D}|\mu, A)p(\mu|A)d\mu \\ p(\mathbf{D}|B) &= \int p(\mathbf{D}|\lambda, B)p(\lambda|B)d\lambda \end{aligned} \tag{2}$$

are the integrated model likelihoods, $p(\mathbf{D}|\mu, A)$ and $p(\mathbf{D}|\lambda, B)$ being the joint likelihood functions of the models and their parameters. Assume for the sake of simplicity a uniform prior distribution for each parameter, e.g.,

$$p(\lambda|B) = \frac{1}{\lambda_{\max} - \lambda_{\min}} \tag{3}$$

for $\lambda_{\min} \leq \lambda \leq \lambda_{\max}$ and let $\lambda_0 \in [\lambda_{\min}, \lambda_{\max}]$ be the maximum likelihood estimate of λ (a derivation of KIC for general forms of the prior distribution is found in work by *Kashyap* [1982] and *Ye et al.* [2008]). Writing $p(\mathbf{D}|\lambda, B) = \exp(\ln p(\mathbf{D}|\lambda, B))$ and expanding $\ln p(\mathbf{D}|\lambda, B)$ in a Taylor series about λ_0 yields, to second order in the estimation error $\lambda - \lambda_0$, a Gaussian probability density function

$$p(\mathbf{D}|\lambda, B) \approx p(\mathbf{D}|\lambda_0, B) \exp\left[-\frac{(\lambda - \lambda_0)^2}{2(\delta\lambda)^2}\right], \tag{4}$$

where

$$(\delta\lambda)^2 = \left[-\frac{\partial^2 \ln p(\mathbf{D}|\lambda, B)}{\partial\lambda^2}\bigg|_{\lambda=\lambda_0}\right]^{-1} \tag{5}$$

is the inverse of observed Fisher information (as opposed to the more common expected Fisher information [*Kass and Raftery*, 1995]), $\delta\lambda$ being a measure of parameter estimation uncertainty. Assuming that the parameter bounds, [$\lambda_{\min}$, $\lambda_{\max}$], do not cause a significant truncation of the Gaussian

[1] Department of Scientific Computing, Florida State University, Tallahassee, Florida, USA.

[2] Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona, USA.

[3] Pacific Northwest National Laboratory, Richland, Washington, USA.

0043-1397/10/2009WR008501$09.00

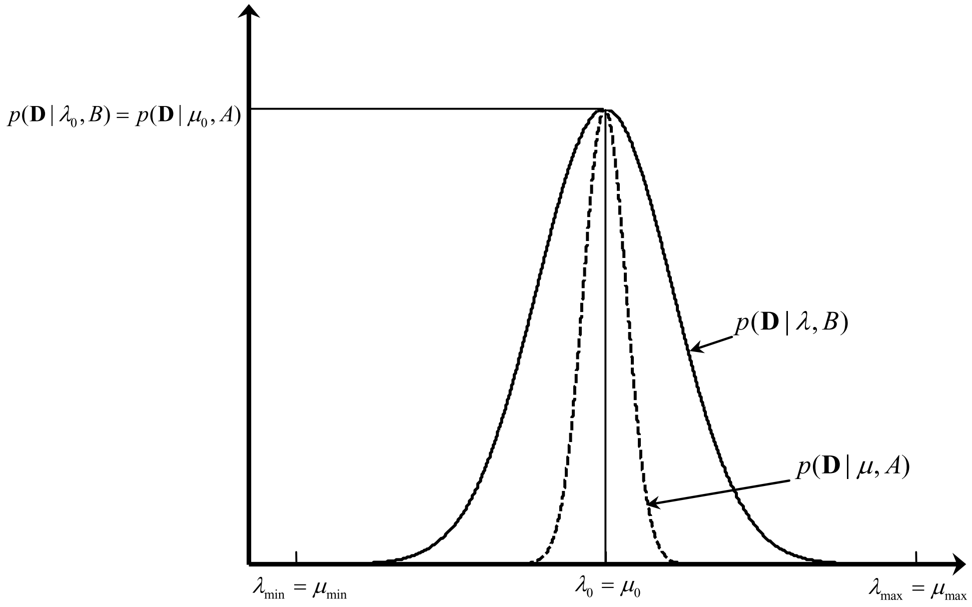

Figure 1. Illustration of the likelihood functions, $p(\mathbf{D}|\lambda, B)$ (solid) and $p(\mathbf{D}|\mu, A)$ (dashed). Models A and B have the same parameter estimates and maximum likelihood values.

function in equation (4), substitution of (3) and (4) into (2) yields the approximation

$$p(\mathbf{D}|B) = \frac{1}{\lambda_{\max} - \lambda_{\min}} \int_{\lambda_{\min}}^{\lambda_{\max}} p(\mathbf{D}|\lambda, B) d\lambda \approx \frac{p(\mathbf{D}|\lambda_0, B)\delta\lambda\sqrt{2\pi}}{\lambda_{\max} - \lambda_{\min}}. \tag{6}$$

This expression is analogous to KIC. Note that KIC derived by *Kashyap* [1982] is the same as equation (5) of *Kass and Raftery* [1995] on the basis of the observed Fisher information matrix. Developing the related approximation for $p(\mathbf{D}|A)$ and $p(\mathbf{D}|B)$ and substituting both into (1) gives

$$\frac{p(A|\mathbf{D})}{p(B|\mathbf{D})} \approx \frac{p(\mathbf{D}|\mu_0, A)}{p(\mathbf{D}|\lambda_0, B)} \frac{\delta\mu}{\delta\lambda} \frac{(\lambda_{\max} - \lambda_{\min})}{(\mu_{\max} - \mu_{\min})} \frac{p(A)}{p(B)}. \tag{7}$$

This equation shows that the relative model preference depends not only on the models' goodness of fit and prior model probabilities but also on the ratios of the Fisher information terms and the prior parameter ranges. The ratio $\delta\mu/\delta\lambda$ between the Fisher terms can thus have a significant influence on model ranking, favoring a model with a larger parameter estimation uncertainty measure.

[4] To better understand the impact of the Fisher information term on model ranking and selection, consider Figure 1 in which the solid curve represents $p(\mathbf{D}|\lambda, B)$ according to (4) with peak at λ_0 and the dashed curve depicts $p(\mathbf{D}|\mu, A)$ with peak at $\mu_0 = \lambda_0$, the two peaks (as well as parameter ranges and prior model probabilities) being identical. In this case, the two models would be associated with identical BIC values, rendering BIC incapable of distinguishing between them. The difference between the two models in Figure 1 is that $\delta\lambda > \delta\mu$; that is, λ_0 is less certain than μ_0 even though the two estimates are the same. As a result, the integrated likelihood of model B is greater than that of model A, i.e., $p(\mathbf{D}|A)/p(\mathbf{D}|B) < 1$. Model B is thus preferred according to (7) even though both models fit the data equally well.

[5] KIC distinguishes between the models in Figure 1 on the basis of the ratio of the Fisher information terms, favoring the model with larger parameter estimation uncertainty. *Ye et al.* [2008] argued that this behavior is desirable because one anticipates a model having large expected information content per observation (and small estimation variance) to exhibit better goodness of fit. All else being equal, if increasing the expected information content of a model fails to improve its performance relative to another model, then selecting a model with greater expected information content would, according to KIC, be unjustified. We consider this logic to be more compelling than that, noted in our introductory paragraph, which underlies the viewpoint of *Tsai and Li* [2008]. Alternatively, one might also view model B in Figure 1 as more robust (than model A), in the sense that the likelihood function near the maximum is less sensitive to deviations of the parameter from its maximum likelihood value. As a result, uncertainty intervals on parameters and predictions are more likely to include the true values. All else being equal, this argues for preference being given to the more robust model.

[6] We close by depicting in Figure 2 an intermediate situation in which the above two models have different maximum likelihood values and corresponding parameter estimates. Without loss of generality, it is again assumed that the models have the same priors. The ratio of the maximum likelihood values is $p(\mathbf{D}|\mu_0, A)/p(\mathbf{D}|\lambda_0, B) = 3/2$, indicating that model A fits the data better than does model B. However, because $\delta\mu/\delta\lambda = 1/3$, the ratio of posterior probability is $p(A|\mathbf{D})/p(B|\mathbf{D}) = 1/2$, indicating that model B is twice as plausible as model A.

[7] While the Fisher information term in KIC provides it with desirable theoretical properties, we do not recommend relying on it blindly. In the example of Figure 2, the fit of

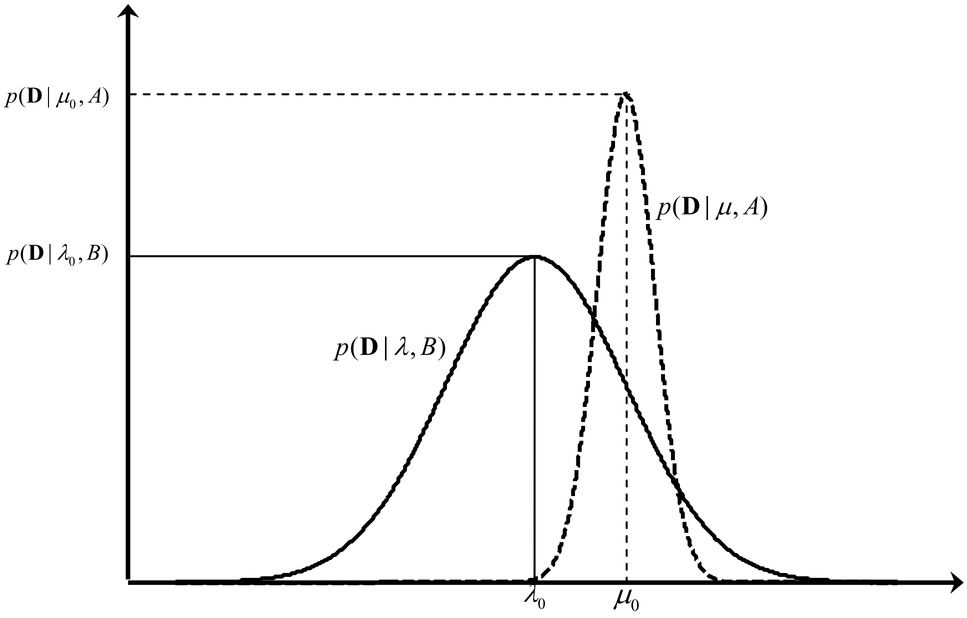

Figure 2. Illustration of the likelihood functions, $p(\mathbf{D}|\lambda, B)$ (solid) and $p(\mathbf{D}|\mu, A)$ (dashed). Models A and B have different parameter estimates and maximum likelihood values.

model B is acceptable. If $\delta\lambda$ were large enough, however, KIC might favor model B even when its goodness of fit is unacceptable. In our own applications of KIC, we have observed such behavior. Model selection criteria such as BIC and KIC are not suitable, however, as the sole means of model evaluation. A model with an unacceptable fit can (and should) be eliminated from consideration by other means, such as inspection of residuals.

[8] Real-world situations are complex and may not conform fully to assumptions behind the derivations of either BIC or KIC. Assumptions that may not be fully satisfied include the validity of disregarding higher-order terms in the derivation of KIC, as we did in (4), which would render the likelihood function more complex and non-Gaussian; ignoring cross correlations between the models and their prior parameter estimates; and/or misrepresenting prior data and parameter statistics. These introduce sufficient ambiguity into the analysis to justify relying on multiple model selection criteria as has become the norm in recent practice. Regardless of such considerations, the inclusion of a Fisher information term in KIC renders it more, not less, suitable than BIC for multimodel assessments on the basis of all but very large data sets. The two criteria often rank models quite differently, our experience suggesting that KIC tends to do so more reliably than BIC [*Carrera and Neuman*, 1986; *Ye et al.*, 2008].

[9] **Acknowledgments.** This work was supported in part by NSF-EAR grant 0911074 and DOE-ERSP grant DE-SC0002687. Work related to this comment by S. P. Neuman was supported in part through a contract between the University of Arizona and Vanderbilt University under the Consortium for Risk Evaluation with Stakeholder Participation (CRESP) III, funded by the U.S. Department of Energy. P. D. Meyer was supported in part by the U.S. Nuclear Regulatory Commission, Office of Nuclear Regulatory Research, under contract N6167. Inspiring discussions with Mary Hill of the U.S Geological Survey and Steffen Mehl of California State University at Chico are appreciated.

References

Carrera, J., and S. P. Neuman (1986), Estimation of aquifer parameters under transient and steady state conditions: 3. Application to synthetic and field data, *Water Resour. Res.*, *22*(2), 228–242, doi:10.1029/WR022i002p00228.

Kashyap, R. L. (1982), Optimal choice of AR and MA parts in autoregressive moving average models, *IEEE Trans. Pattern Anal. Mach. Intell.*, *4*(2), 99–104, doi:10.1109/TPAMI.1982.4767213.

Kass, R. E., and A. E. Raftery (1995), Bayes factors, *J. Am. Stat. Assoc.*, *90*, 773–795, doi:10.2307/2291091.

Schwarz, G. (1978), Estimating dimension of a model, *Ann. Stat.*, *6*(2), 461–464, doi:10.1214/aos/1176344136.

Sivia, D. S., and J. Skilling (2006), *Data Analysis: A Bayesian Tutorial*, 2nd ed., Oxford Univ. Press, Oxford, U. K.

Tsai, F. T.-C., and X. Li (2008), Inverse groundwater modeling for hydraulic conductivity estimation using Bayesian model averaging and variance window, *Water Resour. Res.*, *44*, W09434, doi:10.1029/2007WR006576.

Ye, M., P. D. Meyer, and S. P. Neuman (2008), On model selection criteria in multimodel analysis, *Water Resour. Res.*, *44*, W03428, doi:10.1029/2008WR006803.

D. Lu and M. Ye, Department of Scientific Computing, Florida State University, Tallahassee, FL 32306, USA. (mye@fsu.edu)

P. D. Meyer, Pacific Northwest National Laboratory, Richland, WA 99352, USA.

S. P. Neuman, Department of Hydrology and Water Resources, University of Arizona, Tucson, AZ 85721, USA.