WATER RESOURCES RESEARCH, VOL. 44, W03428, doi:10.1029/2008WR006803, 2008

On model selection criteria in multimodel analysis

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

Received 2 January 2008; accepted 1 February 2008; published 27 March 2008.

[1] Hydrologic systems are open and complex, rendering them prone to multiple conceptualizations and mathematical descriptions. There has been a growing tendency to postulate several alternative hydrologic models for a site and use model selection criteria to (1) rank these models, (2) eliminate some of them, and/or (3) weigh and average predictions and statistics generated by multiple models. This has led to some debate among hydrogeologists about the merits and demerits of common model selection (also known as model discrimination or information) criteria such as *AIC*, *AICc*, *BIC*, and *KIC* and some lack of clarity about the proper interpretation and mathematical representation of each criterion. We examine the model selection literature to find that (1) all published rigorous derivations of *AIC* and *AICc* require that the (true) model having generated the observational data be in the set of candidate models; (2) though *BIC* and *KIC* were originally derived by assuming that such a model is in the set, *BIC* has been rederived by Cavanaugh and Neath (1999) without the need for such an assumption; and (3) *KIC* reduces to *BIC* as the number of observations becomes large relative to the number of adjustable model parameters, implying that it likewise does not require the existence of a true model in the set of alternatives. We explain why *KIC* is the only criterion accounting validly for the likelihood of prior parameter estimates, elucidate the unique role that the Fisher information matrix plays in *KIC*, and demonstrate through an example that it imbues *KIC* with desirable model selection properties not shared by *AIC*, *AICc*, or *BIC*. Our example appears to provide the first comprehensive test of how *AIC*, *AICc*, *BIC*, and *KIC* weigh and rank alternative models in light of the models' predictive performance under cross validation with real hydrologic data.

Citation: 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.

1. Introduction

[2] Hydrologic environments are open and complex, rendering them prone to multiple interpretations and mathematical descriptions regardless of the quantity and quality of available data. This recognition has led to a growing tendency among hydrologists to postulate several alternative hydrologic models for a site and use various criteria to (1) rank these models, (2) eliminate some of them, and/or (3) weigh and average predictions and statistics generated by multiple models (*Neuman* [2003], *Neuman and Wierenga* [2003], *Ye et al.* [2004], *Poeter and Anderson* [2005], *Beven* [2006] and references therein pertaining to GLUE, and *Refsgaard et al.* [2006]). This in turn has brought about a debate among hydrogeologists about the merits and demerits of various model selection (also known as model discrimination or information) criteria such as the information-theoretic criteria *AIC* [*Akaike*, 1974] and *AICc* [*Hurvich and Tsai*, 1989] and the Bayesian criteria *BIC* [*Schwarz*, 1978] and *KIC* [*Kashyap*, 1982]. These criteria discriminate between models based on how closely they reproduce hydrologic observations using maximum likelihood estimates of model parameters (favoring models that reproduce observed behavior most closely) and how many such parameters they contain (penalizing models that contain many). *KIC* additionally considers the likelihood of the parameter estimates in light of their prior values (when such are available) and contains a Fisher information matrix term that as we shall see, imbues it with desirable model selection properties not shared by *AIC*, *AICc*, or *BIC*. Models associated with smaller values of a given criterion are ranked higher than those associated with larger values, the absolute value of the criterion being irrelevant.

[3] Consider a set $\mathbf{M}$ of K alternative models, M_k, $k = 1, 2, \ldots, K$. Then *AIC*, *AICc*, *BIC*, and *KIC* are defined for model M_k as

$$AIC_k = -2\ln\left[L\left(\hat{\boldsymbol{\beta}}_k|\mathbf{z}^*\right)\right] + 2N_k \tag{1}$$

$$AICc_k = -2\ln\left[L\left(\hat{\boldsymbol{\beta}}_k|\mathbf{z}^*\right)\right] + 2N_k + \frac{2N_k(N_k+1)}{N_z - N_k - 1} \tag{2}$$

$$BIC_k = -2\ln\left[L\left(\hat{\boldsymbol{\beta}}_k|\mathbf{z}^*\right)\right] + N_k \ln N_z \tag{3}$$

[1]School of Computational Science and Department of Geological Sciences, Florida State University, Tallahassee, Florida, USA.
[2]Pacific Northwest National Laboratory, Richland, Washington, USA.
[3]Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona, USA.

0043-1397/08/2008WR006803$09.00

$$KIC_k = -2\ln\left[L\left(\hat{\boldsymbol{\beta}}_k|\mathbf{z}^*\right)\right] - 2\ln p\left(\hat{\boldsymbol{\beta}}_k\right) + N_k\ln(N_z/2\boldsymbol{\pi}) + \ln|\bar{\mathbf{F}}_k|, \tag{4}$$

where $\hat{\boldsymbol{\beta}}_k$ is the maximum likelihood (ML) estimate of a vector $\boldsymbol{\beta}_k$ of N_k adjustable parameters (which may include statistical parameters of the calibration data) associated with model M_k; $\mathbf{z}^*$ is an observed vector of N_z random (hydrologic) system state variables $\mathbf{z}$ in space-time, the randomness of which may be inherent (and thus modeled stochastically) or resulting from an additive random error (typically taken to be associated with measurements), common to all K models in the set; $-\ln[\mathrm{L}(\hat{\boldsymbol{\beta}}_k|\mathbf{z}^*)]$ is the minimum of the negative log-likelihood (*NLL*) function $-\ln[\mathrm{L}(\boldsymbol{\beta}_k|\mathbf{z}^*)]$ occurring, by definition, at $\hat{\boldsymbol{\beta}}_k$; $\mathrm{p}(\hat{\boldsymbol{\beta}}_k)$ is the prior probability of $\boldsymbol{\beta}_k$ evaluated at $\hat{\boldsymbol{\beta}}_k$ and $\bar{\mathbf{F}}_k = \mathbf{F}_k/N_z$ is the normalized (by N_z) observed (implicitly conditioned on the observations $\mathbf{z}^*$ and evaluated at the maximum likelihood parameter estimates $\hat{\boldsymbol{\beta}}_k$) Fisher information matrix $\mathbf{F}_k$ having elements [*Kashyap*, 1982]

$$\bar{\mathbf{F}}_{kij} = \frac{1}{N_z}F_{kij} = -\frac{1}{N_z}\frac{\partial^2\ln[L(\boldsymbol{\beta}_k|\mathbf{z}^*)]}{\partial\boldsymbol{\beta}_{ki}\partial\boldsymbol{\beta}_{kj}}\bigg|_{\boldsymbol{\beta}_k=\hat{\boldsymbol{\beta}}_k}. \tag{5}$$

[4] The first term of each criterion, $-2\ln[\mathrm{L}(\hat{\boldsymbol{\beta}}_k|\mathbf{z}^*)]$ measures goodness of fit between predicted and observed system states, $\hat{\mathbf{z}}$ and $\mathbf{z}^*$, respectively; the smaller this term, the better the fit. The terms containing N_k represent measures of model complexity. The criteria thus embody (to various degrees) the principle of parsimony, penalizing models for having a relatively large number of parameters if this does not bring about a corresponding improvement in model fit.

[5] Expressions equivalent to (1)–(4) for the case of Gaussian likelihood functions, which correspond to parameter estimation schemes based on weighted least squares of the kind employed in some hydrologic inverse codes (e.g., PEST [*Doherty*, 2006], UCODE_2005 [*Poeter et al.*, 2005], and MODFLOW2000 [*Hill et al.*, 2000]), are given in Appendix B.

[6] In the ensuing discussion we drop the subscript k from all terms other than M_k and N_k unless required for clarity. We start by noting that *AIC* [*Akaike*, 1974; *Linhart and Zucchini*, 1986; *Bozdogan*, 1987] is based on the Kullback-Leibler (K-L) information, a measure of the discrepancy between a true but unknown representation (model) of reality from which the observations $\mathbf{z}^*$ arise, denoted here as f, and an approximate representation of the same reality, denoted here as the model g having parameters $\boldsymbol{\beta}$. The Kullback-Leibler information can be expressed as [*Akaike*, 1974]

$$I(f;g(\cdot|\boldsymbol{\beta})) = \int \ln\left[\frac{f(\mathbf{z}^*)}{g(\mathbf{z}^*|\boldsymbol{\beta})}\right] f(\mathbf{z}^*)\,d\mathbf{z}^*. \tag{6}$$

[7] *AIC* is an asymptotically unbiased estimator of $E[I(f;g(\cdot|\hat{\boldsymbol{\beta}}))] = \int I(f;g(\cdot|\hat{\boldsymbol{\beta}}))\,f(\mathbf{z}^*)\,d\mathbf{z}^*$ with g evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, the expectation being taken with respect to $f(\mathbf{z}^*)$. As N_z/N_k decreases *AIC* becomes progressively more biased, a property improved upon by *AICc* [*Hurvich and Tsai*, 1989], which constitutes an approximation of $E[I(f;g(\cdot|\hat{\boldsymbol{\beta}}))]$ with better small sample performance. *Burnham and Anderson* [2002] advocate the use of *AICc* when N_z/N_k is less than 40.

[8] *BIC* was derived in a Bayesian context by *Schwarz* [1978] as an asymptotic approximation to a transformation of the posterior probability of a candidate model (*Cavanaugh and Neath* [1999]; other derivations are due to *Akaike* [1977] and *Rissanen* [1978]). *Cavanaugh and Neath* [1999] noted that in the case of large samples, *BIC* favors the model which is a posteriori most probable, i.e., is most plausible in light of the available data. Assuming (as did Schwarz) that the data were generated by a model which belongs to the set of candidate models renders *BIC* consistent in the sense that as the sample size N_z increases relative to N_k, the criterion tends to identify this generating (operating), or true model with probability one.

[9] *KIC* was derived in a Bayesian context by *Kashyap* [1982] as an asymptotic approximation to the model likelihood, i.e., the marginal probability density $p(\mathbf{z}^*|M_k) = \int p(\mathbf{z}^*|\boldsymbol{\beta}, M_k)\,p(\boldsymbol{\beta}|M_k)d\boldsymbol{\beta}$ of the observations conditional on a given model M_k in a set of K such models. Like *BIC*, *KIC* was originally derived on the assumption that the (true) model having generated the data is in the set of candidate models. *KIC* is asymptotic in the sense that the approximation improves as $p(\mathbf{z}^*|\boldsymbol{\beta}, M_k)$ becomes more peaked about $\hat{\boldsymbol{\beta}}$, which will generally occur as the number N_z of observations increases. *KIC* is closely related to the asymptotic Laplace approximation of $p(\mathbf{z}^*|M_k)$ [*Kass and Vaidyanathan*, 1992; *Kass and Raftery*, 1995]. The model likelihood arises when evaluating alternative models using Bayes factors [*Kass and Raftery*, 1995] and model probabilities using Bayesian model averaging (BMA). *Neuman* [2003] proposed using, and *Ye et al.* [2004, 2005] as well as *Meyer et al.* [2007] have implemented, *KIC* in the context of maximum likelihood BMA (MLBMA). It is well established (and we show in Appendix A) that *KIC* reduces asymptotically to *BIC* as N_z becomes large in comparison to N_k (i.e., as $N_z/N_k \rightarrow \infty$). When N_z is not large, *BIC* sometimes prefers models with too few parameters [*Bozdogan and Haughton*, 1998], in which case *KIC* is a more appropriate criterion to use. We note also that when $N_z > 8$ the penalty term $N_k \ln N_z$ in *BIC* is larger than $2N_k$, in which case *BIC* places more emphasis on parsimony than does *AIC* (compare equations (1) and (3)).

[10] Selecting the prior probability in the second term, $-2\ln p(\hat{\boldsymbol{\beta}})$, of *KIC* is considered by *Kass and Wasserman* [1996], and estimating it in the model selection context is discussed by *Kass and Raftery* [1995]. In the special case where prior parameter measurements $\boldsymbol{\beta}^*$ are available, $p(\boldsymbol{\beta})$ may be taken to represent the probability density fuction (pdf) of corresponding measurement errors $(\boldsymbol{\beta}^* - \boldsymbol{\beta})$, and $p(\hat{\boldsymbol{\beta}})$ the pdf of associated residuals $(\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}})$, as done by *Carrera and Neuman* [1986a]. The latter authors proposed absorbing $-2\ln p(\hat{\boldsymbol{\beta}})$ into the leading negative log likelihood term, $-2\ln[\mathrm{L}(\hat{\boldsymbol{\beta}}|\mathbf{z}^*)]$, and included it also in the leading term of *AIC* and *BIC*. As $-2\ln p(\hat{\boldsymbol{\beta}})$ drops out of *KIC* in the asymptotic limit of large N_z/N_k (see Appendix A), this term should be excluded from *BIC*. We are not aware of any theoretical justification for including $-2\ln p(\hat{\boldsymbol{\beta}})$ in *AIC*, AIC_c and/or *BIC* as allowed by *Hill* [1998], *Hill and Tiedeman*

[2007], and *Poeter and Hill* [2007]. In the absence of such justification the presence of $-2\ln p(\hat{\beta})$ in *KIC* appears to be a unique feature of this criterion.

[11] There has been much debate in the model selection literature over the merits and demerits of these and other model selection criteria, without a clear consensus. The primary criticism of *BIC* (and *KIC* by association) is that it assumes one of the models under consideration to be true (representing reality) and that *BIC* and *KIC* are therefore inappropriate in applications where models are, by necessity, simplifications of reality. Our paper is motivated in part by a need to clarify this issue considering its relevance to hydrology (and other fields). In particular, whereas we [*Neuman*, 2003; *Ye et al.*, 2004, 2005; *Meyer et al.*, 2007] have based our approach to multimodel hydrogeologic ranking and inference on the Bayesian criterion *KIC* (which we saw reduces asymptotically to *BIC*), *Poeter and Anderson* [2005] have voiced a preference for the information-theoretic criterion *AICc* (which reduces asymptotically to *AIC*). Citing primarily *Burnham and Anderson* [2002, 2004], *Poeter and Anderson* [2005, p. 604] conclude that

> . . .Approaches based on K-L information view models as approximations of the truth, and assume (1) a true model does not exist and cannot be expected to be in the set of models and (2) as the number of observations increases, one can uncover more details of the system; thus, *AICc* will select more complex models when more observations are available. Alternative model selection criteria (e.g., *BIC*, *HQ*, and *KIC*) seek to identify the true (or quasi-true) model with consistent complexity as the number of observations goes to infinity. These alternatives are based on the assumption that reality can be nearly expressed as a model and that this quasi-true model is in the set. Although these measures may perform similarly in applications, it is unreasonable to assume that they would ever include the true or quasi-true model in the set of alternative ground water models; thus, approaches based on K-L information such as *AICc* are the preferable model ranking and inference criterion.

[12] In the remainder of this paper we address these and related issues in light of the published literature and present a computational example which appears to provide the first comprehensive test of how *AIC*, *AICc*, *BIC*, and *KIC* weigh and rank alternative models in light of the models' predictive performance under cross validation with real hydrologic data. Another application of cross validation to the testing of alternative hydrogeological models (without considering KIC) is provided by *Foglia et al.* [2006, 2007].

2. Underlying Principles

[13] In deriving *AICc* for regression models, *Hurvich and Tsai* [1989, p. 299] state, and their mathematics makes clear, that they "assume. . .the approximating family includes the operating model. This is a strong assumption, but it is also used in the derivation of *AIC* [*Linhart and Zucchini*, 1986, p. 245]," and, we add, in those of *Sawa* [1978] and *Konishi and Kitagawa* [1996]. In deriving *AICc* for autoregressive models, Hurvich and Tsai (p. 305) again "assume that the approximating family includes the operating model". Both regression and autoregressive models are relevant to hydrology, the latter in time series analysis and the former in parameter estimation [*Hill*, 1998; *Hill and Tiedeman*, 2007], both having been employed jointly by *Carrera and Neuman* [1986a, 1986b]. The operating model (the generating model referred to above in the context of *BIC* derivation) is the underlying model that has given rise to the observations [*Hurvich and Tsai*, 1989, pp. 298, 305; *Zucchini*, 2000, p. 42]. *Zucchini* [2000, pp. 52–53] states without ambiguity that the derivation of *AIC* depends on assuming "the operating model belongs to the approximating family".

[14] *Burnham and Anderson* [2002, pp. 362–374] follow *Takeuchi* [1976] in deriving [*Burnham and Anderson*, 2004, p. 270] "an asymptotically unbiased estimator of relative, expected K-L information that applies in general without assuming that model g is true (i.e., without the special conditions underlying Akaike's derivation of AIC)". To derive *AIC* from this general criterion, one must assume [*Burnham and Anderson*, 2002, p. 368] that the true model f is a subset of the approximating family of g, i.e., that "$g = f$ or f is contained within g in the sense of nested models". *Burnham and Anderson* [2002] make the same assumption in deriving the *AICc* result of *Hurvich and Tsai* [1989]. Thus, whereas it may be that [*Burnham and Anderson*, 2004, p. 270] "In practice, one need not assume that the "true model" is in the set of candidates" when using *AIC* or *AICc*, there does not appear to be any rigorous published mathematics to support this theoretically. Contrary to the assertion of *Poeter and Anderson* [2005] that "Approaches based on K-L information" such as *AIC* and *AICc* "assume. . . a true model does not exist and cannot be expected to be in the set of models", the literature demonstrates that the opposite is, in fact, the case.

[15] Though *BIC* and *KIC* were originally derived by assuming that the data-generating model is in the set, *BIC* has been rederived by *Cavanaugh and Neath* [1999] in a Bayesian context which eliminates several restrictive assumptions adopted for this purpose by *Schwarz* [1978]. Most important, their development requires no assumptions regarding the structure of the models entering the set or the way that the data are interrelated statistically. In the words of *Burnham and Anderson* [2002, p. 293], "Cavanaugh and Neath. . .make it clear that the derivation of BIC does not require any assumption about the true model being in the set of models". The assumption by *Kashyap* [1982] in deriving *KIC* that the true model having generated the data be in the set of the candidate models is introduced solely to insure consistency, i.e., that this model is identified by *KIC* with probability 1 in the limit of large sample size. It follows that since *BIC* is an asymptotic limit of *KIC* (Appendix A), the fact that *BIC* does not require the presence of a generating model in the set implies that neither does *KIC* (in which case their consistency property becomes irrelevant). According to *Burnham and Anderson* [2002, p. 294], in the derivation leading to *KIC* (and *BIC*) in our Appendix A, which is analogous to the "heuristic derivation" of *BIC* in section 6.4.1 of their book, "there is no requirement that g be the true model" where g is equivalent to model M_k in our Appendix A. This lays to rest the assertion by *Poeter and Anderson* [2005] that *BIC* and *KIC* require a true or quasi-true model to be in the set of candidate models.

[16] Concerning the interpretation of model probability, although prior model probability, $p(M_k)$, has been interpreted by Kashyap (as well as by *Schwarz* [1978] and *Hoeting et al.* [1999]) as the prior probability that M_k is the true model given that one of the models in M is true, the

Bayesian context is equally compatible with a view of $p(M_k)$ as a subjective prior probability reflecting the analyst's perception about how plausible each alternative model (or a group of models) is relative to other models based on their apparent (qualitative, a priori) consistency with available knowledge and data [*Ye et al.*, 2004, 2005]. The analyst's perception, degree of reasonable belief [*Jeffreys*, 1957], or confidence [*Zio and Apostolakis*, 1996] in a model is ideally based on expert judgment, which *Bredehoeft* [2005] considers to be the basis for conceptual model development. Hence integrating expert judgment into the specification of subjective prior probabilities is a strength rather than a weakness. In the same vein, posterior model probability represents a measure of how plausible each model is relative to all other models within the set in light of its consistency with available knowledge and data following maximum likelihood calibration.

[17] *Burnham and Anderson* [2002, p. 295] state that "there is. . .nothing in the foundation or derivation of BIC that addresses a bias-variance trade-off, and hence addresses parsimony as a feature of BIC model selection. This is not a strike against BIC because this trade-off is a frequentist concept not explicitly invoked in Bayesian statistics. But we are left with no theoretical basis to know what sort of parsimony the BIC model selection procedure has". In our Bayesian view, *BIC* favors a model with just the right sort of parsimony (as well as bias-variance trade-off) that render it, asymptotically, most probable a posteriori of having generated the data (without requiring that this probability go asymptotically to 1). The same applies to *KIC*. We note (and demonstrate by example later) that whereas the term $N_k \ln N_z$ causes *BIC* to prefer less complex models (with fewer parameters) as sample size increases, in *KIC* this tendency is tempered (for all but very large N_z/N_k) by the last term containing a Fisher information matrix as explained in some detail below. This tendency of *KIC* to sometimes prefer more complex models than do either *BIC*, *AIC*, or *AICc* is due to its unique ability to discriminate between models based not only on their number of parameters and sample size but also on the quality of the parameter estimates and the observational data. Whereas *Poeter and Anderson* [2005] may be right in stating that *AICc* will select more complex models when more observations are available, it will do so without regard to the quality of these observations or that of the parameter estimates. This explains why comparisons in the literature between information-theoretic criteria and *BIC* do not generally carry over to *KIC*.

[18] The comparative analysis of model selection criteria of *Burnham and Anderson* [2002, pp. 284–301] is predicated on the notion that (p. 284) "AIC, AICc. . . were motivated by the concept that truth is very complex and that no "true model" exists", which contradicts our finding that all rigorous derivations of *AIC* and *AICc* assume $f \equiv g$ or $f \subset g$. The analysis is limited to hydrologically irrelevant cases in which *BIC* is consistent so that a true or a quasi-true model (the lowest-dimensional among a nested sequence of true models, nested meaning that each model of dimension N_k contains all models of dimension smaller than N_k) is in the set. In this comparison *BIC* fares poorly relative to *AIC* due primarily to the underlying premises. It also appears to perform in a less satisfactory manner than does *AIC* when the candidate models exhibit a tapering effect (smaller and smaller modeling effects are revealed gradually as sample size increases so that the K-L information $I(f;g_k(\cdot|\hat{\boldsymbol{\beta}}))$ in (6) diminishes gradually and monotonically as N_z goes up) of the kind illustrated by *Burnham and Anderson* [2004, pp. 278–279]. We are, however, not aware of any analyses comparing the performance of *AIC*, *AICc*, *BIC*, and/or *KIC* in a comprehensive and convincing manner under assumptions and conditions reflective of actual situations commonly encountered in hydrology. We present such a comparative analysis in this paper.

3. Role of the Fisher Information Term in *KIC*

[19] A major difference between *AIC*, *AICc*, and *BIC* on one hand and *KIC* on the other is the presence of a Fisher information term $\ln|\bar{\mathbf{F}}|$ in equation (4) defining *KIC*. Viewing $\mathbf{z}$ either as a stochastic vector or as a deterministic vector of "true" system states, and $\mathbf{z}^*$ either as a sample of a stochastic $\mathbf{z}$ or as a vector of "true" $\mathbf{z}$ values corrupted by measurement errors, renders the vector $\mathbf{z}^*$ random. A convenient way to explain the role that $\ln|\bar{\mathbf{F}}|$ plays in *KIC* is to consider the random vector $\boldsymbol{\varepsilon}^* = \mathbf{z}^* - \mathbf{z}$ as having a Gaussian distribution with zero mean and a covariance matrix $\boldsymbol{C}_z$ (observations that are not Gaussian can be transformed into a Gaussian form). Writing $\boldsymbol{C}_z = \mathbf{T}\mathbf{T}^T$ where $\mathbf{T}$ is a lower or an upper triangular matrix (other such decompositions are possible) allows writing $\boldsymbol{\varepsilon}^* = \mathbf{T}\boldsymbol{\zeta}^*$ where $\boldsymbol{\zeta}^*$ is a random vector of mutually uncorrelated components having zero mean and unit variance. Then the weighted sum of squares $\boldsymbol{\varepsilon}^{*T}\mathbf{C}_z^{-1}\boldsymbol{\varepsilon}^*$ entering into a Gaussian likelihood

$$L(\boldsymbol{\beta}|\mathbf{z}^*) = p(\mathbf{z}^*|\boldsymbol{\beta}) = (2\pi)^{-N_z/2}|\mathbf{C}_z|^{-1/2}\exp\left(-\frac{1}{2}\boldsymbol{\varepsilon}^{*T}\mathbf{C}_z^{-1}\boldsymbol{\varepsilon}^*\right) \tag{7}$$

can be replaced by the simple sum of squares $\boldsymbol{\zeta}^{*T}\boldsymbol{\zeta}^*$, the components of which, being Gaussian, are independent and identically distributed. It follows that each element of $\bar{\mathbf{F}}$ can be written as a sum of terms,

$$\bar{F}_{ij} = -\frac{1}{N_z}\sum_{n=1}^{N_z}\frac{\partial^2 \ln[p(\zeta_n^*|\boldsymbol{\beta})]}{\partial\beta_i\partial\beta_j}. \tag{8}$$

Correspondingly, the law of large numbers implies that $\bar{F}_{ij}$ converges to the expectation of a single term,

$$\bar{F}_{ij} \to -E\left[\frac{\partial^2 \ln\left[p\left(\zeta_1^*|\boldsymbol{\beta}\right)\right]}{\partial\beta_i\partial\beta_j}\right] = \left\langle F_{1ij}\right\rangle \quad \text{as} \quad N_z \to \infty, \tag{9}$$

where $\langle\mathbf{F}_1\rangle$ is the expected Fisher information associated with a single observation.

[20] If the expected information content per observation as expressed by $\ln|\bar{\mathbf{F}}|$ does not vary from model to model, then *KIC* rewards models in proportion to the quality of the fit they provide between predictions and a given set of observations (as measured by the negative log likelihood function $-2\ln[\mathrm{L}(\hat{\boldsymbol{\beta}}|\mathbf{z}^*)]$, common to all four criteria) and between posterior parameter estimates and their prior values (as measured by $-2\ln p(\hat{\boldsymbol{\beta}})$, which we saw is not included in *BIC* and does not appear to be a valid component of either *AIC* or *AICc*). At the same time, *KIC* penalizes models in proportion to their number of parameters, N_k, through the

term $N_k \ln (N_z/2\pi)$, which differs from corresponding terms in *BIC*, *AIC*, and *AICc* and depends to a lesser extent on sample size, N_z, than does $N_k \ln N_z$ in *BIC*. Yet among models having an equal number of parameters, and equal values of $-2\ln[L(\hat{\boldsymbol{\beta}}|\mathbf{z}^*)] - 2\ln p(\hat{\boldsymbol{\beta}})$, *KIC* favors models with relatively small expected information content per observation. This is consistent with the criterion of sufficiency according to which [*Fisher*, 1922, p. 316], "the statistic chosen should summarize the whole of the relevant information supplied by the sample". Since *KIC* considers the parameter estimates to be normally distributed (see Appendix A), and a Gaussian distribution of estimates with a larger variance (smaller information content) per unit sample contains a similar distribution with a smaller variance (larger information content) per unit sample, the former summarizes the multimodel population (the whole) more fully than does the latter and is therefore chosen. Looking at this from a different but related angle, one anticipates a model having large expected information content per observation (and small estimation variance) to exhibit improved performance (exhibit better combined fits to observations and prior parameters) and vice versa. 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.

[21] Among models having different numbers of parameters with equal combined fits to observations and prior parameters, *KIC* balances parsimony as expressed by $N_k \ln(Nz/2\pi)$ with expected information content per observation as expressed by $\ln|\bar{\mathbf{F}}|$. If under these circumstances $N_k \ln(N_z/2\pi)$ of model M_k exceeds $N_l \ln(N_z/2\pi)$ of model M_l while $N_l \ln(N_z/2\pi) + \ln|\bar{\mathbf{F}}_l|$ exceeds $N_k \ln(N_z/2\pi) + \ln|\bar{\mathbf{F}}_k|$, then *KIC* selects the less parsimonious model M_k over model M_l because the latter, though more parsimonious, contains a greater amount of expected information per observation than does the former. The inclusion of $\ln|\bar{\mathbf{F}}|$ in *KIC* is thus seen to imbue it with a unique ability to discriminate between models based not only on how well they fit observations and how many parameters they contain as do *AIC*, *AICc*, and *BIC* but also on how close are the posterior parameter estimates to their prior values and how much expected information is contained, on average, in each observation.

[22] By virtue of (5) the Fisher information term of *KIC* in (4) can be written as

$$\ln|\bar{\mathbf{F}}| = \ln\left|\frac{1}{N_z}\mathbf{F}\right| = \ln\left(N_z^{-N_k}|\mathbf{F}|\right) = -N_k \ln N_z + \ln|\mathbf{F}|, \quad (10)$$

where $\mathbf{F}$ is the observed Fisher information matrix. Substituting (10) into (4) gives

$$KIC = -2\ln\left[L\left(\hat{\boldsymbol{\beta}}|\mathbf{z}^*\right)\right] - 2\ln p(\hat{\boldsymbol{\beta}}) - N_k \ln 2\pi + \ln|\mathbf{F}|, \quad (11)$$

which is the Laplace approximation of model likelihood given by *Kass and Vaidyanathan* [1992] and *Kass and Raftery* [1995]. The expected Fisher information matrix having elements

$$\langle F_{ij}\rangle = -E\left[\frac{\partial^2 \ln[L(\boldsymbol{\beta}|\mathbf{z}^*)]}{\partial\beta_i \partial\beta_j}\right]\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} \quad (12)$$

is often interchanged with $\mathbf{F}$. Substituting $\langle\mathbf{F}\rangle$ for $\mathbf{F}$ in (4) or (11) increases the order of the error in the approximation [*Kass and Raftery*, 1995]. The inverse expected Fisher information matrix is the Cramer-Rao (lower) bound of the covariance matrix, $\boldsymbol{\Sigma}$, of the ML parameter estimates, $\hat{\boldsymbol{\beta}}$ [*Papoulis*, 1991]. It is common to set

$$\boldsymbol{\Sigma} = \langle\mathbf{F}\rangle^{-1} \quad (13)$$

as an approximation [e.g., *Carrera and Neuman*, 1986a]. Assuming equivalence of $\langle\mathbf{F}\rangle$ and $\mathbf{F}$ [*Efron and Hinkley*, 1978; *Kass and Raftery*, 1995] and substituting (13) into (11) gives

$$KIC = -2\ln[L\left(\hat{\boldsymbol{\beta}}|\mathbf{z}^*\right)] - 2\ln p(\hat{\boldsymbol{\beta}}) - N_k \ln 2\pi - \ln|\boldsymbol{\Sigma}|. \quad (14)$$

When prior information about the hydrologic parameters is unavailable, the term $-2\ln p(\hat{\boldsymbol{\beta}})$ drops out. An equivalent expression corresponding to parameter estimation schemes based on weighted least squares, of the kind employed in some hydrologic inverse codes (e.g., PEST [*Doherty*, 2006], UCODE_2005 [*Poeter et al.*, 2005], and MODFLOW2000 [*Hill et al.*, 2000]), is given in Appendix B.

[23] Past applications of model selection criteria in the hydrologic literature have not always used expressions for *KIC* consistent with those presented here. Although *Carrera and Neuman* [1986a] presented the correct expression for *KIC*, in computing *KIC* values for their example *Carrera and Neuman* [1986b] apparently substituted $\langle\mathbf{F}\rangle$ for $\bar{\mathbf{F}}$ in (4); their application of (14) was missing the term $-N_k \ln N_z$. *Hernandez et al.* [2006] likewise left out this term. *Poeter and Anderson* [2005] appear to have left out two terms, $-P_k \ln\hat{\sigma}^2_{ML}$ and $-P_k \ln N_z$ where P_k is the number of hydrologic model parameters and $\hat{\sigma}^2_{ML}$ is a ML estimate of a nominal observation error variance (all defined in Appendix B), due to an erroneous specification of the Fisher information matrix.

4. Comparative Analysis of Model Selection Criteria

[24] We present below what appears to be the first comparative analysis of how well *AIC*, *AICc*, *BIC*, and *KIC* discriminate between models given information about the models' ability to render hydrologic predictions of measured quantities. Our analysis concerns the geostatistical characterization of log air permeability ($\log_{10}k$) in unsaturated fractured tuff at the Apache Leap Research Site (ALRS) in central Arizona. It is based on 184 pneumatic injection tests in 1-m-long segments of six vertical and inclined (at 45°) boreholes at the site. Seven alternative variogram models were fitted to these data by *Ye et al.* [2004]: power (*Pow0*), exponential without a drift (*Exp0*), exponential with a linear drift (*Exp1*), exponential with a quadratic drift (*Exp2*), spherical without a drift (*Sph0*), spherical with a linear drift (*Sph1*), and spherical with a quadratic drift (*Sph2*). The authors used an adjoint state maximum likelihood cross validation (ASMLCV) method due to *Samper and Neuman* [1989] in conjunction with universal kriging [*Deutsch and Journel*, 1998] and general-

Table 1. Predictive Log Score of Each $\log_{10}k$ Variogram Model and of Model-Averaged Results Based on *AIC*, *AICc*, *BIC*, and *KIC* in Cross Validations Against Measured $\log_{10}k$ Values in Boreholes V2, X2, Y2, Y3, Z2, and W2A at the Apache Leap Research Site

	Predictive Log Scores of Individual Models						
Borehole	*Pow0*	*Exp0*	*Exp1*	*Exp2*	*Sph0*	*Sph1*	*Sph2*
V2	21.63	23.48	23.27	25.64	24.88	31.59	32.05
X2	29.68	27.05	27.88	27.91	32.60	41.08	41.11
Y2	27.28	29.33	24.13	25.06	30.65	32.76	32.85
Y3	55.79	59.11	39.58	45.27	59.62	53.24	53.31
Z2	37.97	38.70	48.35	92.32	39.18	57.78	58.03
W2A	32.30	33.79	40.61	24.96	35.50	33.68	33.68
Average	34.11	35.24	33.97	40.19	37.07	41.69	41.84

	Log Scores of Model-Averaged Predictions			
Borehole	*AIC*	*AICc*	*BIC*	*KIC*
V2	21.64	21.64	21.63	21.79
X2	28.04	28.07	29.30	27.64
Y2	24.46	24.33	26.01	24.25
Y3	40.73	40.57	45.43	41.82
Z2	44.09	43.08	38.02	40.41
W2A	32.15	32.26	32.31	32.48
Average	31.85	31.66	32.12	31.40

ized least squares [*Neuman and Jacobson*, 1984] to obtain unbiased ML estimates of variogram parameters and drift coefficients for each of the seven models. They used ASMLCV rather than the ML estimator to associate each measurement with a variance σ^2 and computed the observed Fisher information matrix directly [*Ye et al.*, 2004, equation (9)] rather than on the basis of a Jacobian matrix.

[25] In a manner patterned after *Ye et al.* [2004] we cross validate below each of the seven variogram models by checking their ability to predict log permeability in each of the six boreholes based solely on measured values in the remaining five boreholes. We translate the values of *AIC*, *AICc*, *BIC*, and *KIC* associated with each model into a posterior model weight, $p(M_k|\mathbf{z}^*)$, which in the case of *BIC* and *KIC* represents posterior model probability. The posterior model weights are computed according to

$$p(M_k|\mathbf{z}^*) = \frac{\exp\left(-\frac{1}{2}\Delta IC_k\right)p(M_k)}{\sum_{l=1}^{7}\exp\left(-\frac{1}{2}\Delta IC_l\right)p(M_l)}, \qquad (15)$$

where $\Delta IC_k = IC_k - IC_{\min}$, IC_k being any of the four model selection or information criteria and $IC_{\min} = \min_k\{IC_k\}$ its minimum over all seven models. To render our analysis neutral, we set the prior probability $p(M_k)$ of each of the seven models, M_k, $k = 1, 2, \ldots, 7$, equal to 1/7. Next, we characterize the predictive capabilities of each model in terms of its predictive log score (a measure of information lost upon eliminating from consideration some of the data for purposes of cross validation; calculation of the log score is based on equations (20)–(22) of *Ye et al.* [2004]) and predictive coverage (percent of eliminated data falling within the 90% prediction interval). The lower the predictive log score and the higher the predictive coverage of a given model, the better its predictive capabilities. We compute similar measures of performance for predictions obtained upon averaging those of all seven individual models using the corresponding values of $p(M_k|\mathbf{z}^*)$, based on each of the four information criteria, as weights. Details of our cross-validation and model averaging procedures are given by *Ye et al.* [2004].

[26] Table 1 lists the predictive log score results corresponding to individual variogram models and to model averaged predictions based on *AIC*, *AICc*, *BIC*, and *KIC* obtained through cross validation against measured $\log_{10}k$ values in boreholes V2, X2, Y2, Y3, Z2, and W2A. Predictive log scores obtained through model averaging were smaller than values obtained using individual models, regardless of which model selection criterion was used (Table 1). Differences in predictive log scores obtained using individual models were considerably larger than differences in model-average results between model selection criteria. These results suggest that in terms of predictive performance, choosing to use model-average predictions (instead of predictions from a single model) is much more important than the choice of a particular model selection criterion.

[27] Predictive coverage results, shown in Table 2, are similar to the log score results. Predictive coverage values obtained through model averaging were generally larger than those obtained using individual models except in the case of *AIC* and *AICc*, which yielded smaller coverage values than did models *Pow0* and *Sph0* (Table 2). Differences between predictive coverage values obtained using individual models were considerably larger than differences in model-average results between model selection criteria. Nonetheless, Tables 1 and 2 indicate that the overall model-averaged performance of *KIC* (as measured by an average log score of 31.40 and an average predictive coverage of 87.46%) was superior to that of the other three criteria, the least satisfactory log score being that of *BIC* and the lowest predictive coverage that of *AICc*.

Table 2. Predictive Coverage (%) of Each $\log_{10}k$ Variogram Model and of Model-Averaged Results Based on *AIC*, *AICc*, *BIC*, and *KIC* in Cross Validations Against Measured $\log_{10}k$ Values in Boreholes V2, X2, Y2, Y3, Z2, and W2A at the Apache Leap Research Site

	Predictive Coverage of Individual Models, %						
Borehole	*Pow0*	*Exp0*	*Exp1*	*Exp2*	*Sph0*	*Sph1*	*Sph2*
V2	100.00	60.00	80.00	70.00	80.00	75.00	70.00
X2	90.00	90.00	93.33	90.00	93.33	96.67	90.00
Y2	89.29	92.86	85.71	78.57	92.86	92.86	92.86
Y3	57.50	60.00	77.50	77.50	62.50	72.50	72.50
Z2	82.14	82.14	82.14	50.00	82.14	78.57	53.57
W2A	100.00	100.00	83.78	51.35	100.00	86.49	83.78
Average	86.49	80.83	83.75	69.57	85.14	83.68	77.12

	Coverage of Model-Averaged Predictions, %			
Borehole	*AIC*	*AICc*	*BIC*	*KIC*
V2	100.00	100.00	100.00	100.00
X2	93.33	93.33	90.00	93.33
Y2	89.29	89.29	92.86	89.29
Y3	77.50	70.00	57.50	60.00
Z2	50.00	50.00	82.14	82.14
W2A	100.00	100.00	100.00	100.00
Average	85.02	83.77	87.08	87.46

Table 3. Posterior Model Weights (%) and Corresponding Ranks Assigned to Each Variogram Model by *AIC*, *AICc*, *BIC*, and *KIC* Compared With Log Score and Predictive Coverage Values and Ranks Based on Actual Model Performance Under Cross Validation[a]

Model	*Pow0*	*Exp0*	*Exp1*	*Exp2*	*Sph0*	*Sph1*	*Sph2*	*SAD*
AIC	13.65%	0.17%	50.60%	34.17%	0.00%	0.91%	0.50%	...
Rank	3	6	1	2	7	4	5	16
AICc	19.44%	0.24%	58.76%	20.21%	0.00%	1.06%	0.30%	...
Rank	3	6	1	2	7	4	5	16
BIC	98.21%	1.19%	0.59%	0.00%	0.00%	0.01%	0.00%	...
Rank	1	2	3	6	5	4	7	6
KIC	35.30%	26.58%	37.61%	0.00%	0.00%	0.51%	0.00%	...
Rank	2	3	1	6	5	4	7	6
Log score	34.11	35.24	33.97	40.19	37.07	41.69	41.84	...
Rank	2	3	1	5	4	6	7	...
Coverage	86.49%	80.83%	83.75%	69.57%	85.14%	83.68%	77.12%	...
Rank	1	5	3	7	2	4	6	...
SPR	3	8	4	12	6	10	13	...
Rank	1	4	2	6	3	5	7	...

[a]*SPR* is the sum of ranks based on log score and predictive coverage; the bottom row is the rank based on *SPR*. *SAD* is the sum of absolute differences between the rank assigned to each model by a model selection criterion and the rank based on *SPR*.

[28] Table 3 compares posterior model weights (probabilities in the case of *BIC* and *KIC*) (%) and corresponding ranks assigned to each variogram model by *AIC*, *AICc*, *BIC*, and *KIC* with log score and predictive coverage values and ranks based on actual model performance under cross validation. Table 3 also lists the sum of model ranks based on log score and predictive coverage (SPR), ranks based on SPR, and the sum of absolute differences (SAD) between the rank assigned to each model by a given information criterion and the rank based on SPR. Table 3 shows that (1) *BIC* and *KIC* clearly identified the two best performing (*Pow0* and *Exp1*) and the two worst performing (*Sph2* and *Exp2*) models, (2) *AIC* and *AICc* ranked the second worst performing model (*Exp2*) as second best, and (3) in terms of the overall measure SAD, *BIC*, and *KIC* ranked the models much more consistently with their order of performance (as measured by the cross validation) than did *AIC* and *AICc*.

[29] Additional details and insights concerning our cross-validation results are offered in Appendix C.

5. Summary of Key Findings

[30] The following is a summary of our key findings:

[31] 1. All published rigorous derivations of the information-theoretic model selection criteria *AIC* and *AICc* require that the (true) model, which has generated the observational data, be in the set of models being analyzed.

[32] 2. Though the Bayesian model selection criteria *BIC* and *KIC* were originally derived by assuming that the true model is in the set of candidate models, *BIC* has subsequently been rederived without the need for such an assumption.

[33] 3. *KIC* reduces to *BIC* as the number of observations becomes large relative to the number of adjustable model parameters, implying that it likewise does not require the existence of a true model in the set of alternatives.

[34] 4. If a true model is in the set, *BIC* and *KIC* select with probability one the true model as sample size increases, a consistency property not shared by *AIC* and *AICc*.

[35] 5. Published comparisons between *BIC* and *AIC* (none considers *KIC* and few consider *AICc*) tend to rely on the consistency of *BIC*, which does not apply when a true model is not in the set, as is usually the case in hydrology.

[36] 6. *KIC* is the only criterion accounting validly for the likelihood of prior parameter estimates.

[37] 7. The presence of a Fisher information term in *KIC* imbues it with desirable model selection properties not shared by *AIC*, *AICc*, or *BIC*. *KIC* sometimes prefers more complex models than do other criteria due to its unique ability to discriminate between models based not only on 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. Whereas *AICc* may select more complex models as the number of observations increases, it does so without regard to such quality considerations. *BIC* and *KIC* are sufficiently different from each other to render published comparisons between information-theoretic criteria and *BIC* inapplicable to *KIC*.

[38] 8. Our computational example appears to provide the first comparative analysis of how well *AIC*, *AICc*, *BIC*, and *KIC* discriminate between models given information about the models' ability to render hydrologic predictions of measured quantities under cross validation. In the example, variogram parameters are estimated using adjoint-state and nonlinear regression methods, similar to those employed widely in estimating groundwater flow and transport parameters. In our example, predictions obtained through weighted model averaging were generally better than those generated by individual models regardless of which model selection criterion was used to assign the weights. Differences among predictions obtained through model averaging based on various model selection criteria were much smaller than among those generated by individual models. While this suggests that choosing to use model averaging is a much more important decision than choosing a particular model selection criterion, the overall model-averaged performance of *KIC* was better than that of any other criterion tested, the least satisfactory measures of performance being associated with *BIC* (highest log score) and *AICc* (lowest percent of predictive coverage).

[39] 9. Whereas weights assigned to the various models in our example by *AIC* and *AICc* tended to be similar,

posterior probabilities assigned to the models by *BIC* and *KIC* tended to differ markedly from each other and from the *AIC*/*AICc* weights. *KIC* is strongly and uniquely influenced by its Fisher information term, which increased sharply with the number of model parameters within each of two nested sets of exponential and spherical variogram models we considered.

[40] 10. Whereas *BIC* and *KIC* clearly identified the two best performing (*Pow0* and *Exp1*) and the two worst performing (*Sph2* and *Exp2*) models among seven variogram models in our example, *AIC* and *AICc* ranked the second worst performing model (*Exp2*) as second best. *BIC* and *KIC* ranked the models much more consistently with their actual order of performance than did *AIC* and *AICc*.

[41] 11. Only one of the seven sample variograms obtained upon cross validating data from one of six boreholes (Y3) in our example appeared to represent a stationary random field (see Appendix C). *KIC* was the only information criterion that recognized this by favoring a stationary variogram model (*Exp0*) in this case.

[42] 12. Predictions obtained using variogram model *Exp2* when cross validating data from one of six boreholes (Z2) in our example were very different from those generated by *Pow0*, *Exp0*, and *Exp1* (see Appendix C), reflecting what appeared to be an excessive fit to noisy data (due to a relatively large number of adjustable parameters). Whereas *Exp2* was ranked in this case second to last among all seven models by *BIC* and *KIC*, it was ranked second best by *AIC* and *AICc*. Similar rankings were reflected in model-averaged results, which in the case of *AIC* and *AICc* were dominated by *Exp2*. This exemplifies the known tendency of *AIC* and *AICc* to favor models which exhibit a closer fit to data than these data warrant (i.e., the tendency to over fit, which some confuse with accuracy).

Appendix A

[43] Following *Kashyap* [1982], for K competing (linear or nonlinear, Gaussian or non-Gaussian) models M_k, the posterior probability $p(M_k|\mathbf{z}^*)$ of M_k conditioned on the observations $\mathbf{z}^*$ is expressed according to Bayes' rule as

$$p(M_k|\mathbf{z}^*) = \frac{p(\mathbf{z}^*|M_k)p(M_k)}{p(\mathbf{z}^*)} = C_k p(\mathbf{z}^*|M_k). \tag{A1}$$

The likelihood function $p(\mathbf{z}^*|M_k)$ of M_k is given by

$$p(\mathbf{z}^*|M_k) = \int p(\mathbf{z}^*|\boldsymbol{\beta}_k, M_k)\, p(\boldsymbol{\beta}_k|M_k)\, d\boldsymbol{\beta}_k, \tag{A2}$$

where $\boldsymbol{\beta}_k$ (of dimension N_k) is the vector of parameters associated with M_k. Let $\hat{\boldsymbol{\beta}}_k$ be the maximum likelihood estimates of $\boldsymbol{\beta}_k$ based on observations $\mathbf{z}^*$ and the likelihood function $L(\boldsymbol{\beta}_k|\mathbf{z}^*, M_k) = p(\mathbf{z}^*|\boldsymbol{\beta}_k, M_k)$. Expressing the latter as $\exp(\ln p(\mathbf{z}^*|\boldsymbol{\beta}_k, M_k))$, expanding $\ln p(\mathbf{z}^*|\boldsymbol{\beta}_k, M_k)$ and $p(\boldsymbol{\beta}_k|M_k)$ in a Taylor series about $\hat{\boldsymbol{\beta}}_k$ and ignoring higher-order terms in the parameter estimation error $(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)$ gives

$$p(\mathbf{z}^*|M_k) = p\left(\mathbf{z}^*\middle|\hat{\boldsymbol{\beta}}_k, M_k\right) p\left(\hat{\boldsymbol{\beta}}_k|M_k\right) \int \exp\left(-\frac{1}{2}\left(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right)^T N_z \bar{\mathbf{F}}_k(\mathbf{z}^*|\boldsymbol{\beta}_k, M_k)\left(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right)\right) d\boldsymbol{\beta}_k, \tag{A3}$$

where the normalized (by N_z) observed Fisher information matrix $\bar{\mathbf{F}}_k(\mathbf{z}^*|\boldsymbol{\theta}_k, M_k)$ is defined as

$$\bar{F}_{kij}(\mathbf{z}^*|\boldsymbol{\beta}_k, M_k) = -\frac{1}{N_z}\frac{\partial^2 \ln p(\mathbf{z}^*|\boldsymbol{\beta}_k, M_k)}{\partial\beta_i\, \partial\beta_j}\bigg|_{\boldsymbol{\beta}_k = \hat{\boldsymbol{\beta}}_k}, \tag{A4}$$

N_z being the dimension of $\mathbf{z}^*$. Near $\boldsymbol{\beta}_k$ the estimation error $(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k)$ is close to Gaussian so that

$$1 = (2\pi)^{-\frac{N_k}{2}}\left|N_z \bar{\mathbf{F}}_k\left(\mathbf{z}^*\middle|\hat{\boldsymbol{\beta}}_k, M_k\right)\right|^{\frac{1}{2}} \int \exp\left(-\frac{1}{2}\left(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right)^T N_z \bar{\mathbf{F}}_k(\mathbf{z}^*|\boldsymbol{\beta}_k, M_k)\left(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_k\right)\right) d\boldsymbol{\beta}_k. \tag{A5}$$

Substituting (A5) into (A3) yields

$$p(\mathbf{z}^*|M_k) = p\left(\mathbf{z}^*\middle|\hat{\boldsymbol{\beta}}_k, M_k\right) p\left(\hat{\boldsymbol{\beta}}_k|M_k\right) \times (2\pi/N_z)^{\frac{N_k}{2}}\left|\mathbf{F}_k\left(\mathbf{z}^*\middle|\hat{\boldsymbol{\beta}}_k, M_k\right)\right|^{-\frac{1}{2}}. \tag{A6}$$

Taking natural logarithm and multiplying by -2 leads directly to (4). Since $-2\ln[L(\hat{\boldsymbol{\beta}}_k|\mathbf{z}^*)]$ in (4) is of order N_z it follows from (5) that $\bar{F}_{ij}$ is of order 1 and $\ln|\bar{\mathbf{F}}|$ is of order $\ln N_k$ while $-2\ln p(\hat{\boldsymbol{\beta}}_k)$ is of order N_k as is $-N_k \ln(2\pi)$. In the limit of a large sample N_z relative to N_k ($N_z/N_k \rightarrow \infty$), terms of order N_k and $\ln N_k$ can be disregarded in comparison to terms of order N_z and *KIC* reduces to *BIC* in (3). A similar derivation of *KIC* and discussion of the relationship between *KIC* and *BIC* are given by *Kass and Raftery* [1995], *Neath and Cavanaugh* [1997], and *Congdon* [2001].

Appendix B

[44] The covariance $\mathbf{C}_z$ of the observation errors $\boldsymbol{\varepsilon}^* = \mathbf{z}^* - \mathbf{z}$ is sometimes expressed as [*Carrera and Neuman*, 1986a; *Hill*, 1998] $\mathbf{C}_z = \boldsymbol{\sigma}^2\, \boldsymbol{\omega}^{-1}$ where ω is a known weight matrix and σ^2 is a known or unknown nominal error variance. When prior estimates of hydrologic parameters are available, they can be incorporated into $\mathbf{z}^*$ as proposed by *Carrera and Neuman* [1986a] and *Cooley* [1983] for deterministic models and by *Hernandez et al.* [2006] for stochastic ensemble moment models; in both cases the prior parameter estimates are taken to be uncorrelated with the observational system state errors, yielding a block diagonal covariance matrix $\mathbf{C}_z$. Substituting $\mathbf{C}_z = \boldsymbol{\sigma}^2\, \boldsymbol{\omega}^{-1}$ into (7) gives

$$L(\boldsymbol{\beta}_k|\mathbf{z}^*) = (2\pi)^{-N_z/2}\left|\boldsymbol{\sigma}^2\boldsymbol{\omega}^{-1}\right|^{-1/2} \exp\left(-\frac{1}{2}\frac{\boldsymbol{\varepsilon}^{*T}\boldsymbol{\omega}\boldsymbol{\varepsilon}^*}{\sigma^2}\right), \tag{B1}$$

which in turn can be rewritten as

$$-2\ln[L(\boldsymbol{\beta}_k|\mathbf{z}^*)] = N_z \ln(2\pi) + N_z \ln\sigma^2 + \ln|\boldsymbol{\omega}^{-1}| + \frac{\boldsymbol{\varepsilon}^{*T}\boldsymbol{\omega}\boldsymbol{\varepsilon}^*}{\sigma^2}. \tag{B2}$$

[45] If σ^2 is known, $\boldsymbol{\beta}_k = \boldsymbol{\theta}_k$ where θ_k is a vector of P_k hydrologic model parameters so that $N_k = P_k$. In this case minimizing the above negative log likelihood criterion

Table C1. Values of Negative Log-Likelihood (*NLL*), *AIC*, *AICc*, *BIC*, $\ln|\bar{\mathbf{F}}|$, *KIC*, and Relative Weights or (in the Case of *BIC* and *KIC*) Posterior Probabilities $p(M_k|\mathbf{z}^*)$ Assigned by Each Information Criterion to Each Variogram Model M_k, $k = 1,\ldots 7$, During Cross Validation of Measured Log Permeabilities in Three (X2, Y3, and Z2) of Six Boreholes at the Apache Leap Research Site[a]

Model	*Pow0*	*Exp0*	*Exp1*	*Exp2*	*Sph0*	*Sph1*	*Sph2*
N_k	2	2	6	12	2	6	12
				X2 (N_z=154)			
NLL	297.13	304.00	285.74	276.67	318.14	291.69	281.56
AIC	301.13	308.00	297.74	300.67	322.14	303.69	305.56
$p(M_k\|\mathbf{z}^*)$	12.32%	0.40%	67.02%	15.50%	0.00%	3.42%	1.34%
AICc	301.21	308.08	298.31	302.88	322.22	304.26	307.77
$p(M_k\|\mathbf{z}^*)$	16.74%	0.54%	71.21%	7.25%	0.00%	3.63%	0.63%
BIC	307.20	314.07	315.96	337.11	328.21	321.91	342.00
$p(M_k\|\mathbf{z}^*)$	95.66%	3.08%	1.20%	0.00%	0.00%	0.06%	0.00%
$\ln\|\bar{\mathbf{F}}\|$	10.32	1.75	5.96	43.25	3.91	8.31	44.76
KIC	313.85	312.15	310.90	358.31	328.45	319.19	364.71
$p(M_k\|\mathbf{z}^*)$	12.85%	30.07%	56.18%	0.00%	0.01%	0.89%	0.00%
				Y3 (N_z=144)			
NLL	273.40	277.75	265.01	252.59	291.43	273.70	264.70
AIC	277.40	281.75	277.01	276.59	295.43	285.70	288.70
$p(M_k\|\mathbf{z}^*)$	26.04%	2.96%	31.55%	38.94%	0.00%	0.41%	0.09%
AICc	277.48	281.83	277.63	278.97	295.51	286.31	291.08
$p(M_k\|\mathbf{z}^*)$	39.50%	4.49%	36.75%	18.73%	0.00%	0.48%	0.04%
BIC	283.34	287.68	294.83	312.23	301.37	303.51	324.34
$p(M_k\|\mathbf{z}^*)$	89.51%	10.18%	0.29%	0.00%	0.01%	0.00%	0.00%
$\ln\|\bar{\mathbf{F}}\|$	10.28	2.06	6.25	44.47	4.41	6.60	40.71
KIC	289.94	286.07	290.06	334.65	302.10	299.08	342.99
$p(M_k\|\mathbf{z}^*)$	11.25%	78.01%	10.60%	0.00%	0.03%	0.12%	0.00%
				Z2 (N_z=156)			
NLL	264.23	272.95	250.59	232.31	291.12	262.47	236.42
AIC	268.23	276.95	262.59	256.31	295.12	274.47	260.42
$p(M_k\|\mathbf{z}^*)$	0.22%	0.00%	3.69%	85.14%	0.00%	0.01%	10.93%
AICc	268.31	277.03	263.15	258.49	295.20	275.03	262.60
$p(M_k\|\mathbf{z}^*)$	0.60%	0.01%	7.90%	81.07%	0.00%	0.02%	10.41%
BIC	274.33	283.05	280.89	292.91	301.22	292.77	297.01
$p(M_k\|\mathbf{z}^*)$	95.18%	1.21%	3.58%	0.01%	0.00%	0.01%	0.00%
$\ln\|\bar{\mathbf{F}}\|$	10.83	2.46	6.32	43.54	4.98	10.23	44.65
KIC	281.48	281.83	276.18	314.40	302.52	291.97	319.61
$p(M_k\|\mathbf{z}^*)$	6.24%	5.24%	88.49%	0.00%	0.00%	0.03%	0.00%

[a] N_z is the number of data used to estimate N_k model parameters.

(maximizing the likelihood L) is equivalent to minimizing the generalized (or weighted if $\boldsymbol{\omega}$ is diagonal as given by *Hill* [1998]) least squares (LS) criterion $\mathbf{e}^T\boldsymbol{\omega}\boldsymbol{e}$. This yields a ML and, equivalently, a LS estimate $\hat{\boldsymbol{\beta}}_k = \hat{\boldsymbol{\theta}}_k$ of $\boldsymbol{\beta}_k = \boldsymbol{\theta}_k$. If $\boldsymbol{\sigma}^2$ is unknown, one must estimate via ML an extended (hydrologic and statistical) parameter vector $\boldsymbol{\beta}_k = \{\boldsymbol{\theta}_k, \sigma^2\}$ having dimension $N_k = P_k + 1$; LS estimation would not yield an ML estimate of σ^2 as required for the evaluation of *AIC*, *AICc*, *BIC*, and *KIC*. Since σ^2 is often difficult to evaluate a priori, we focus below on this latter case.

[46] Considering that $\boldsymbol{\theta}_k$ and σ^2 are mutually independent, it is possible to obtain a ML estimate $\hat{\boldsymbol{\theta}}_k$ of $\boldsymbol{\theta}_k$ without knowing σ^2 by setting $-2\partial\ln[L(\boldsymbol{\theta}_k, \sigma^2|\mathbf{z}^*)]/\partial\boldsymbol{\theta}_k = 0$. Since $N_z\ln(2\pi)$, $N_z \ln \sigma^2$, and $\ln|\boldsymbol{\omega}^{-1}|$ in (B2) are independent of $\boldsymbol{\theta}_k$, this is equivalent to minimizing the (generalized or weighted) sum of squared residuals $\mathbf{e}^T\boldsymbol{\omega}\mathbf{e}$ where $\mathbf{e} = \hat{\mathbf{z}} - \mathbf{z}^*$ and $\hat{\mathbf{z}}$ is the computed value of $\mathbf{z}$, yielding an estimate $\hat{\boldsymbol{\theta}}_k$ that is at once ML and LS. One can then estimate σ^2 a posteriori by setting $-2\partial\ln[L(\hat{\boldsymbol{\theta}}_k, \sigma^2|\mathbf{z}^*)]/\partial\sigma^2 = 0$, yielding the ML estimate [*Seber and Wild*, 1989; *Seber and Lee*, 2003; *Carrera and Neuman*, 1986a]

$$\hat{\sigma}^2_{ML} = \left.\frac{\mathbf{e}^T\boldsymbol{\omega}\mathbf{e}}{N_z}\right|_{\boldsymbol{\theta}_k = \hat{\boldsymbol{\theta}}_k}. \tag{B3}$$

As this ML estimate is biased, it is sometimes replaced by the unbiased LS estimator [*Seber and Wild*, 1989; *Seber and Lee*, 2003; *Hill*, 1998]

$$\hat{\boldsymbol{\sigma}}^2_{LS} = \left.\frac{\mathbf{e}^T\boldsymbol{\omega}\mathbf{e}}{N_z - P_k}\right|_{\boldsymbol{\theta}_k = \hat{\boldsymbol{\theta}}_k}. \tag{B4}$$

When $N_z \gg P_k$, the difference between (B3) and (B4) is negligible, but in the more general ML context within which *AIC*, *AICc*, *BIC*, and *KIC* are defined, especially in the hydrologically significant case where N_z is not much larger than *Pk* (for which *AICc* and *KIC* are the most appropriate), there does not appear to be a theoretical justification for replacing $\hat{\sigma}^2_{ML}$ with $\hat{\sigma}^2_{LS}$, even if the former is biased. Accordingly, substituting (B3) into (B2) yields

$$-2\ln\left[L\left(\hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{\sigma}}^2_{ML}|\mathbf{z}^*\right)\right] = N_z\ln(2\pi) + N_z\ln\hat{\boldsymbol{\sigma}}^2_{ML} + \ln|\boldsymbol{\omega}^{-1}| + N_z. \tag{B5}$$

Since $N_z\ln(2\pi)$, $\ln|\boldsymbol{\omega}^{-1}|$ and N_z are constant (independent of the choice of model or parameters), they do not affect

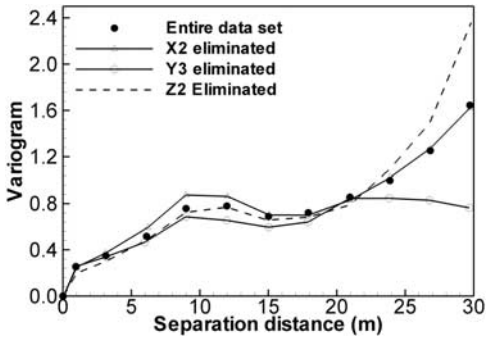

Figure C1. Omnidirectional sample variograms of all log permeability (m^2) data and of partial data sets obtained upon eliminating those measured in boreholes X2, Y3, and Z2.

model selection or relative weighting. Substituting (B5) with $N_k = P_k + 1$ into (1)–(3) and dropping all constants leads to

$$AIC_k = N_z \ln \hat{\sigma}^2_{ML} + 2P_k \tag{B6}$$

$$AICc_k = N_z \ln \hat{\sigma}^2_{ML} + 2P_k + \frac{2N_k(N_k+1)}{N_z - N_k - 1} \tag{B7}$$

$$BIC_k = N_z \ln \hat{\sigma}^2_{ML} + P_k \ln N_z. \tag{B8}$$

Substituting (B5) into (14) and dropping all constants gives

$$KIC_k = N_z \ln \hat{\sigma}^2_{ML} - 2 \ln p(\hat{\boldsymbol{\theta}}_k) - P_k \ln 2\pi - \ln |\boldsymbol{\Sigma}_k| \tag{B9}$$

with $-2\ln p(\hat{\boldsymbol{\theta}})$ dropping out when prior information about the hydrologic parameters is not available. In this latter case [*Carrera and Neuman*, 1986a]

$$\boldsymbol{\Sigma}_k^{-1} = \frac{1}{\hat{\sigma}^2_{ML}} \mathbf{J}_k^T \boldsymbol{\omega} \mathbf{J}_k, \tag{B10}$$

where $\mathbf{J}_k$ is the Jacobian (sensitivity) matrix having elements $J_{kij} = -\partial \hat{z}_i / \partial \theta_{kj}|_{\boldsymbol{\theta}_k = \hat{\boldsymbol{\theta}}_k}$ that linearizes nonlinear groundwater models around the maximum likelihood (or least squares) parameter estimates. Hence

$$\ln|\boldsymbol{\Sigma}_k| = P_k \ln \hat{\sigma}^2_{ML} - \ln\left|\mathbf{J}_k^T \boldsymbol{\omega} \mathbf{J}_k\right| \tag{B11}$$

$$KIC_k = (N_z - P_k) \ln \hat{\sigma}^2_{ML} - P_k \ln 2\pi + \ln\left|\mathbf{J}_k^T \boldsymbol{\omega} \mathbf{J}_k\right|. \tag{B12}$$

Appendix C

[47] Here we provide further information about the comparative analysis of information criteria discussed in the text. Table C1 lists values of *NLL* (negative log-likelihood), *AIC*, *AICc*, *BIC*, $\ln|\bar{\mathbf{F}}|$, *KIC*, and relative weights or (in the case of *BIC* and *KIC*) posterior probabilities $p(M_k|\mathbf{z}^*)$ assigned by each information criterion to each variogram model M_k, $k = 1, 2, \ldots, 7$, during the cross validation of measured log permeabilities in three of the six boreholes (X2, Y3, and Z2) at the Apache Leap Research Site (ALRS). As expected, *NLL* generally decreased as the number of variogram parameters (*NP*) increased, most notably so among nested sets of models belonging to the exponential and spherical families. Whereas weights assigned to the various models by *AIC* and *AICc* tended to be similar, posterior probabilities assigned to the models by *BIC* and *KIC* tended to differ markedly from each other and from the *AIC*/*AICc* weights. Differences between posterior probabilities assigned to the models by *BIC* and *KIC* were due in part to differences between corresponding $\ln|\bar{\mathbf{F}}|$ values which increased sharply with N_k within each nested set of exponential and spherical models, affecting *KIC* but not *BIC*. Hence comparisons in the literature between *AIC*, *AICc*, and *BIC* do not generally extend to *KIC*.

[48] Figure C1 depicts omnidirectional sample variograms of all 184 log permeability (m^2) data and of partial data sets obtained upon eliminating those measured in boreholes X2, Y3, and Z2. This and Figure 11 of *Ye et al.* [2004] indicate that only one of the sample variograms, obtained upon eliminating data from borehole Y3, appeared to represent a stationary field. Table C1 shows *KIC* was the only information criterion to recognize this by favoring a stationary variogram model (*Exp0*) in this case. We note here that the number of data pairs associated with each point

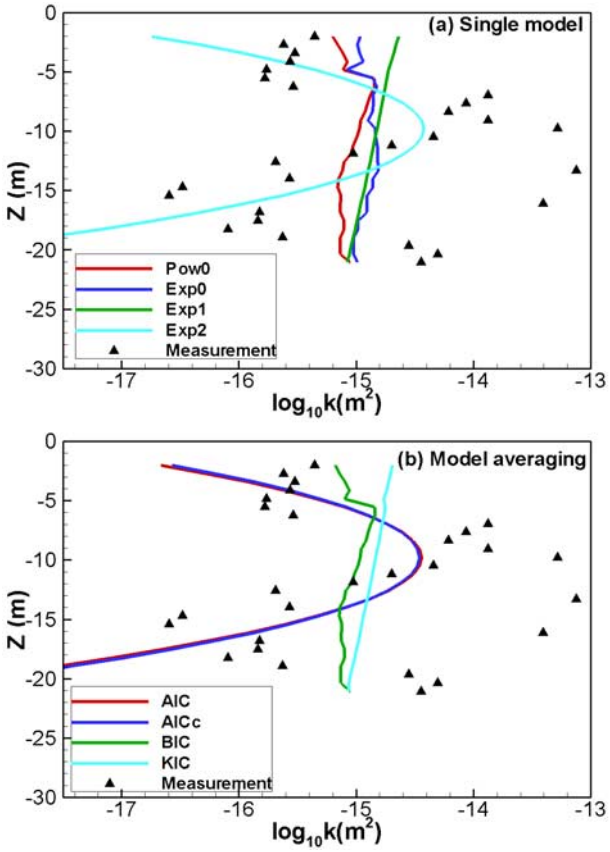

Figure C2. Predicted log permeabilities (m^2) using (a) *Pow0*, *Exp0*, *Exp1*, and *Exp2* and (b) model averaging based on *AIC*, *AICc*, *BIC*, and *KIC* versus measured values obtained during cross validation of data from borehole Z2.

on each sample variogram, including those associated with cross validating the remaining three boreholes, exceeded several hundred at all lags.

[49] Figure C2a compares predicted log permeabilities using variogram models *Pow0*, *Exp0*, *Exp1*, and *Exp2* with measured values obtained during cross validation of data from borehole Z2 (spherical model results were inferior and therefore are not shown). None of the four models captured fully the spatial variability of log permeabilities at the site. Predictions obtained using variogram model *Exp2* were very different from those obtained using *Pow0*, *Exp0*, and *Exp1*, reflecting what appears to be an excessive fit to noisy data (due to the relatively large number of adjustable parameters in model *Exp2*). It is thus noteworthy that *Exp2* was ranked second to last among all seven models by *BIC* and *KIC* but second best by *AIC* and *AICc* (Table 3). These rankings are reflected in model-averaged results in Figure C2b, which, in the case of *AIC* and *AICc*, were dominated by *Exp2*. This exemplifies the known tendency of *AIC* and *AICc* to favor models that exhibit a closer fit to data than these data warrant (i.e., the tendency to over fit, which some confuse with accuracy).

[50] **Acknowledgments.** This research was supported by the U.S. Nuclear Regulatory Commission, Office of Nuclear Regulatory Research, under contracts JCN Y6465 and N6167 with Pacific Northwest National Laboratory. The first author was additionally supported by the FYAP program of Florida State University. The University of Arizona component of the work was supported in part through a contract with Vanderbilt University under the Consortium for Risk Evaluation with Stakeholder Participation (CRESP) III, funded by the U.S. Department of Energy. We thank Mary Hill and Eileen Poeter for their comments.

References

Akaike, H. (1974), A new look at statistical model identification, *IEEE Trans. Autom. Control*, *AC-19*, 716–722.

Akaike, H. (1977), On entropy maximization principle, in *Applications of Statistics*, edited by P. R. Krishnaiah, pp. 27–41, North-Holland, Amsterdam.

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

Bozdogan, H. (1987), Model selection and Akaike's information criterion (AIC): The general theory and its analytical expansions, *Psychometrika*, *52*(3), 345–370.

Bozdogan, H., and D. M. A. Haughton (1998), Informational complexity criteria for regression models, *Comput. Stat. Data Anal.*, *28*, 51–76.

Bredehoeft, J. (2005), The conceptualization model problem-surprise, *Hydrogeol. J.*, *13*, 37–46.

Burnham, K. P., and D. R. Anderson (2002), *Model Selection and Multiple Model Inference: A Practical Information-Theoretical Approach*, 2nd ed., Springer, New York.

Burnham, K. P., and D. R. Anderson (2004), Multimodel inference—Understanding AIC and BIC in model selection, *Sociol. Methods Res.*, *33*(2), 261–304.

Carrera, J., and S. P. Neuman (1986a), Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information, *Water Resour. Res.*, *22*(2), 199–210.

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

Cavanaugh, J. E., and A. A. Neath (1999), Generalizing the derivation of the Schwarz information criterion, *Commun. Stat. Theory Methods*, *28*, 49–66.

Congdon, P. (2001), *Bayesian Statistical Modeling*, John Wiley, Hoboken, N. J.

Cooley, R. L. (1983), Incorporation of prior information on parameters into nonlinear regression groundwater flow models: 2. Applications, *Water Resour. Res.*, *19*(3), 662–676.

Deutsch, C. V., and A. G. Journel (1998), *GSLIB: Geostatistical Software Library and User's Guide*, 2nd ed., Oxford Univ. Press, New York.

Doherty, J. (2006), PEST Model-Independent Parameter Estimation, V10.1, Papadopulos. S. S., Inc., Bethesda, Md. (Available at http://www.sspa.com/pest/)

Efron, B., and D. V. Hinkley (1978), Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information, *Biometrika*, *65*(3), 457–482.

Fisher, R. A. (1922), On the mathematical foundations of theoretical statistics, *Philos. Trans. R. Soc., Ser. A.*, *222*, 309–368.

Foglia, L., S. W. Mehl, M. C. Hill, and P. Burlando (2006), Use of cross-validation to analyze predictive capabilities of alternative groundwater models, in *Proceeding of the Conference MODFLOW and More 2007: Managing Ground Water Systems*, edited by E. P. Poeter et al., pp. 243–247, Colo. School of Mines, Golden.

Foglia, L., S. W. Mehl, M. C. Hill, P. Perona, and P. Burlando (2007), Testing alternative ground water models using cross validation and other methods, *Ground Water*, *45*(5), 627–641.

Hernandez, A. F., S. P. Neuman, A. Guadagnini, and J. Carrera (2006), Inverse stochastic moment analysis of steady state flow in randomly heterogeneous media, *Water Resour. Res.*, *42*, W05425, doi:10.1029/2005WR004449.

Hill, M. C. (1998), Methods and guidelines for effective model calibration, *U.S. Geol. Surv. Water Resour. Invest. Rep., 98-4005*, 90 pp.

Hill, M. C., and C. R. Tiedeman (2007), *Effective Methods and Guidelines for Groundwater Model Calibration, Including Analysis of Data, Sensitivities, Predictions, and Uncertainty*, John Wiley, Hoboken, N. J.

Hill, M. C., E. R. Banta, A. W. Harbaugh, and E. R. Anderman (2000), MODFLOW-2000, the U. S. Geological Survey modular groundwater model—User guide to the observation, sensitivity, and parameter estimation processes and three post-processing programs, *U.S. Geol. Surv. Open File Rep., 00-184*, 209 pp.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999), Bayesian model averaging: A tutorial, *Stat. Sci.*, *14*(4), 382–417.

Hurvich, C. M., and C.-L. Tsai (1989), Regression and time series model selection in small sample, *Biometrika*, *76*(2), 99–104.

Jeffreys, H. (1957), *Scientific Inference*, 2nd ed., Cambridge Univ. Press., New York.

Kashyap, R. L. (1982), Optimal choice of AR and MA parts in autoregressive moving average models, *IEEE Trans. Pattern Anal. Machine Intell.*, *4*(2), 99–104.

Kass, R. E., and A. E. Raftery (1995), Bayesian factor, *J. Am. Stat. Assoc.*, *90*, 773–795.

Kass, R. E., and S. K. Vaidyanathan (1992), Approximate Bayes factors and orthogonal parameters, with application to testing equality of two binomial proportions, *J. R. Stat. Soc., Ser. B*, *54*(1), 129–144.

Kass, R. E., and L. Wasserman (1996), The selection of prior distributions by formal rules, *J. Am. Stat. Assoc.*, *91*(435), 1343–1370.

Konishi, S., and G. Kitagawa (1996), Generalised information criterion in model selection, *Biometrika*, *83*, 875–890.

Linhart, H., and W. Zucchini (1986), *Model Selection*, John Wiley, Hoboken, N. J.

Meyer, P. D., M. Ye, M. L. Rockhold, S. P. Neuman, and K. J. Cantrell (2007), Combined estimation of hydrogeologic conceptual model, parameter, and scenario uncertainty with application to uranium transport at the Hanford Site 300 area, *NUREG/CR-6940 (PNNL-16396)*, U.S. Nucl. Regul. Comm., Washington, D. C.

Neath, A. A., and J. E. Cavanaugh (1997), Regression and time series model selection using variants of the Schwarz information criterion, *Commun. Stat. Theory Methods*, *26*, 559–580.

Neuman, S. P. (2003), Maximum likelihood Bayesian averaging of alternative conceptual-mathematical models, *Stochast. Environ. Res. Risk Assess.*, *17*(5), 291–305, doi:10.1007/s00477-003-0151-7.

Neuman, S. P., and E. A. Jacobson (1984), Analysis of nonintrinsic spatial variability by residual kriging with application to regional groundwater levels, *Math. Geol.*, *16*, 491–521.

Neuman, S. P., and P. J. Wierenga (2003), A comprehensive strategy of hydrogeologic modeling and uncertainty analysis for nuclear facilities and sites, *NUREG/CR-6805*, U.S. Nucl. Regul. Comm., Washington, D. C.

Papoulis, A. (1991), *Probability, Random Variables, and Stochastic Processes*, 3rd ed., McGraw-Hill, New York.

Poeter, E. P., and D. A. Anderson (2005), Multimodel ranking and inference in ground water modeling, *Ground Water*, *43*(4), 597–605.

Poeter, E. P., and M. C. Hill (2007), MMA, a computer code for multimodel analysis, *U.S. Geol. Surv. Tech. Methods, TM6-E3*, 113 pp.

Poeter, E. P., M. C. Hill, E. R. Banta, S. Mehl, and S. Christensen (2005), UCODE_2005 and six other computer codes for universal sensitivity analysis, calibration, and uncertainty evaluation, *U.S. Geol. Surv. Tech. Methods, 6-A11*, 283 pp.

Refsgaard, J. C., J. P. van der Sluijs, J. Brown, and P. van der Keur (2006), A framework for dealing with uncertainty due to model structure error, *Adv. Water. Resour.*, *29*, 1586–1597.

Rissanen, J. (1978), Modeling by shortest data description, *Automatica*, *14*, 465–471.

Samper, F. J., and S. P. Neuman (1989), Estimation of spatial covariance structures by adjoint state maximum likelihood cross validation: 1, *Theory, Water Resour. Res.*, *25*(3), 351–362.

Sawa, T. (1978), Information criteria for discriminating among alternative regression models, *Econometrica*, *46*, 1273–1291.

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Stat.*, *6*(2), 461–464.

Seber, G. A. F., and A. J. Lee (2003), *Linear Regression Analysis*, 2nd ed., John Wiley, Hoboken, N. J.

Seber, G. A. F., and C. J. Wild (1989), *Nonlinear Regression*, John Wiley, Hoboken, N. J.

Takeuchi, K. (1976), Distribution of informational statistics and a criterion of model fitting (in Japanese), *Suri Kagaku*, *153*, 12–18.

Ye, M., S. P. Neuman, and P. D. Meyer (2004), Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff, *Water Resour. Res.*, *40*, W05113, doi:10.1029/2003WR002557.

Ye, M., S. P. Neuman, P. D. Meyer, and K. F. Pohlmann (2005), Sensitivity analysis and assessment of prior model probabilities in MLBMA with application to unsaturated fractured tuff, *Water Resour. Res.*, *41*, W12429, doi:10.1029/2005WR004260.

Zio, E., and G. E. Apostolakis (1996), Two methods for the structured assessment of model uncertainty by experts in performance assessments of radioactive waste repositories, *Reliability Eng. Syst. Safety*, *54*, 225–241.

Zucchini, W. (2000), A introduction to model selection, *J. Math. Psychol.*, *44*, 41–61.

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

S. P. Neuman, Department of Hydrology, University of Arizona, 1133 E. James E. Rogers Way, Tucson, AZ 85721, USA. (neuman@hwr.arizona.edu)

M. Ye, School of Computational Science, Florida State University, Tallahassee, FL 32306, USA.