

Multimodel Ranking and Inference in Ground Water Modeling

by Eileen Poeter¹ and David Anderson²

Abstract

Uncertainty of hydrogeologic conditions makes it important to consider alternative plausible models in an effort to evaluate the character of a ground water system, maintain parsimony, and make predictions with reasonable definition of their uncertainty. When multiple models are considered, data collection and analysis focus on evaluation of which model(s) is(are) most supported by the data. Generally, more than one model provides a similar acceptable fit to the observations; thus, inference should be made from multiple models. Kullback-Leibler (K-L) information provides a rigorous foundation for model inference that is simple to compute, is easy to interpret, selects parsimonious models, and provides a more realistic measure of precision than evaluation of any one model or evaluation based on other commonly referenced model selection criteria. These alternative criteria strive to identify the true (or quasi-true) model, assume it is represented by one of the models in the set, and given their preference for parsimony regardless of the available number of observations the selected model may be underfit. This is in sharp contrast to the K-L information approach, where models are considered to be approximations to reality, and it is expected that more details of the system will be revealed when more data are available. We provide a simple, computer-generated example to illustrate the procedure for multimodel inference based on K-L information and present arguments, based on statistical underpinnings that have been overlooked with time, that its theoretical basis renders it preferable to other approaches.

Introduction

Sparse subsurface data cause us to be uncertain of the exact nature of ground water system structure and components. Consequently, it is a best, although not always customary, practice to evaluate multiple models of a ground water system before making predictions of system behavior. Alternative models include variations in the structure of hydrogeologic units, boundary conditions, and parameter fields. Each alternative model must be calibrated (i.e., parameter values adjusted to obtain the best fit to the observed data, e.g., using nonlinear least squares) before models can be compared (Poeter and Hill 1997). The

advent of high-speed computing and robust inversion algorithms makes calibration of multiple models feasible.

We often find that prediction uncertainty is larger across the range of potential models than that which arises from the misfit and insensitivity of any one optimized model, even to the extent that confidence intervals on predictions from some of the models may not include the values predicted by others. This raises the question of whether to select the best model and use those predictions and confidence intervals for decision and design or to weight all the models and calculate model-averaged predictions and intervals. If one model is clearly superior to the rest, it is reasonable to use that model for prediction, but its uncertainty should be evaluated using the entire set of candidate models. If one model is not clearly superior, then it is reasonable to weight all predictions. If the alternative models yield substantially different results for the prediction of interest such that a reasonable decision is untenable, then additional data should be collected to develop better models.

A more representative model of ground water system behavior (1) exhibits no consistent spatial or temporal

¹Corresponding author: International Ground Water Modeling Center, Department of Geology and Geological Engineering, Colorado School of Mines, 1516 Illinois Street, Golden, CO 80401; (303) 273-3829; fax (303) 384-2037; epoeter@mines.edu

²Applied Information Company, 707 Breakwater Drive, Fort Collins, CO 80525; (970) 229-0255; fax (970) 229-0255; aicanderson1@comcast.net

Received July 2004, accepted October 2004.

Copyright © 2005 National Ground Water Association.

pattern in the weighted residuals; (2) results in reasonable estimated parameter values (e.g., hydraulic conductivity of gravel is higher than that of silt and falls within the range of values that might be expected for gravels); and (3) has better fit statistics for the same data while maintaining parsimony (i.e., balancing the bias vs. variance trade-off or the trade-off between underfitting and overfitting). There is a general agreement that considerable mental effort, training, and experience are required to define a set of reasonable models (Bredehoeft 2003; Neuman and Wierenga 2003). However, the profession has not agreed upon a procedure for ranking or weighting models (Carrera and Neuman 1986; Neuman and Wierenga 2003; Ye et al. 2004).

We have several objectives. First, we call attention to the famous geologist Chamberlin's (1890) call for "multiple working hypotheses" as a strategy for rapid advances in understanding applied and theoretical problems. Each hypothesis or conceptualization is represented by a mathematical model, which gives rigor to the procedure, then data collection and analysis focus on which model is the best, that is, most supported by the data. Second, we introduce a simple and effective approach for the selection of a best model: one that balances underfitting and overfitting (i.e., maintains parsimony). Third, we provide an effective method for making formal multimodel inference, including prediction, from all models in a candidate set. Finally, we present a computer-generated example to illustrate the method, and we comment on alternative approaches.

Model Ranking and Inference from the Best Model

Multiple Working Hypotheses

Ideally, understanding in science comes from strict experimentation. Here, causation can be identified and interactions can be explored. In most cases, an array of practical considerations prevent experimentation in ground water studies. At the opposite extreme are studies that are merely descriptive. Here, progress in understanding is slow and risky. Lack of causation, and other issues, makes this a relatively poor approach. Between these extremes lie studies that can be termed "observational," where inference is model based. One attempts to extract the information in the data using a model. Many ground water problems are in this observational category, and inferences are inherently model based, thus the need for multimodel inference in ground water modeling.

Given a well-defined ground water problem, with extensive thoughtful consideration a hydrologist can conceptualize R hypotheses concerning the system and the questions to be asked. R might range from two to three to perhaps a few dozen or even 100s in cases where statistical techniques are used to generate realizations. Given a good set of data, hypotheses, and models, an investigator can ask, "which hypothesis is most supported by the data?" This is the model selection problem and the heart of Chamberlin's strategic approach. Model selection is a fundamental part of the data analysis. Approaches to optimal inference for one model and data set are known

(e.g., least squares or maximum likelihood methods). The central issue is "which model to use?"

Model Selection

A large effort has been spent on a coherent theory of model selection over the past 30 years. We will not review this material in detail as it is covered in a number of books (e.g., Linhart and Zucchini 1986; McQuarrie and Tsai 1998; Burnham and Anderson 2002), research monographs (e.g., Sakamoto et al. 1986), and hundreds of journal papers (e.g., deLeeuw 1992). Instead, we briefly outline the approach we recommend.

The starting point for effective model selection theory is Kullback-Leibler (K-L) information, $I(f,g)$ (Kullback and Leibler 1951). This is interpreted as the information, I , lost when full truth, f , is approximated by a model, g . Given a set of candidate models g_i , one might compute K-L information for each of the R models and select the one that minimizes information loss—that is, minimize $I(f,g)$ across models. This is a compelling approach. However, for ground water models, K-L information cannot be computed because the truth and the optimal effective parameters (e.g., hydraulic conductivities, boundary heads, and fluxes) are not known (Anderson 2003).

Akaike (1973, 1974) provided a simple way to estimate expected K-L information, based on a bias-corrected, maximized log-likelihood value. This was a major breakthrough (Parzen et al. 1998). Soon thereafter, better approximations to the bias were derived (Sugiura 1978; Hurvich and Tsai 1989, 1994) and the result, of relevance here, is an estimator Akaike Information Criterion (AICc) of twice the expected K-L information loss

$$\text{AICc} = n \log(\sigma^2) + 2k + \left(\frac{2k(k+1)}{n-k-1} \right) \quad (1)$$

where σ^2 is the estimated residual variance, n is the number of observations, and k is the number of estimated parameters for the model. Here, the estimator of $\sigma^2 = \text{WSSR}/n$, where WSSR is the weighted sum of squared residuals. The second term accounts for first-order bias, and the third term accounts for second-order bias resulting from a small number of observations. This is a precise mathematical derivation, with the third term depending on the assumed distribution of residuals, in this case, normally distributed error. Accounting for second-order bias is important when $n/k < 40$, which is typical of ground water models. The aforementioned expression applies to analyses undertaken by a least squares approach; similar expressions are available for those using maximum likelihood procedures (Akaike 1973). AICc is computed for each of the models; the model with the lowest AICc value is the best model, and the remaining models are ranked from best to worst, with increasing AICc values.

As parameters are added to a model, accuracy and variance increase (fit improves, while uncertainty increases). Use of AICc selects models with a balance between accuracy and variance; this is the principle of parsimony. Prediction can be further improved by basing inference on all the models in the set (multimodel inference, as discussed later).

Delta Values

Calculation of the AICc values can be posed so as to retain or omit values that are constant across models (e.g., multinomial coefficients) and are affected by the number of observations; thus, it is essential to compute and use simple differences

$$\Delta_i = \text{AICc}_i - \text{AICc}_{\min} \quad (2)$$

for each model, i , in the set of R models, where AICc_{\min} is the minimum AICc value of all the models in the set. These values are on an information scale ($-\log[\text{probability}]$), free from constants and sample size issues. A Δ_i represents the information loss of model i relative to the best model. As discussed by Burnham and Anderson (2002, p. 70–72 and particularly 78), models with $\Delta_i < 2$ are very good models, while models with $4 < \Delta_i < 7$ have less empirical support. In most cases, models with Δ_i greater than ~ 10 can be dismissed from further consideration.

Model Probabilities

Simple transformation yields model probabilities or Akaike weights (also referred to as posterior model probabilities)

$$w_i = \frac{\exp^{-0.5\Delta_i}}{\sum_{j=1}^R \exp^{-0.5\Delta_j}} \quad (3)$$

where w_i is the weight of evidence in favor of model i being the best model in the sense of minimum K-L information loss. These weights are also useful in multimodel inference as discussed later.

Evidence Ratios

It is convenient to take ratios of the model probabilities for models i and j as w_i/w_j and call these evidence ratios. These are most useful when i is the best model and j is another model of interest because they can be used to make statements such as “there is ‘ w_i/w_j ’ times more evidence supporting the best model.”

Example Problem

Our goal is to illustrate model evaluation first by calibrating a set of simple (coarse versions of the “truth”) ground water models of a synthetic (known) system (as defined by a generating model), then making multimodel inference of predictions. The alternative models used for the example are simplistic relative to models of field sites using only zonation variations generated by a geostatistical simulator. We do not offer this as a desired approach to model development, only as a method for generating models to demonstrate the procedure. Each coarse model is calibrated by weighted least squares nonlinear regression under the initial pumping condition using 20 head observations and 1 base flow observation. Then, we rank and determine weights for the models. In the predictive stage, additional pumping is simulated at another location and head is predicted at 20 locations (offset from the

calibration data locations), while two flows are also predicted. In a subsequent section, we illustrate multimodel inference of the predicted heads and flow rates and compare them to the known predictions simulated by the generating model.

Synthetic Model

A two-dimensional, unconfined steady-state system is synthesized with a model domain 5000 m in the east-west direction and 3000 m north-south direction (Figure 1). The aquifer is assigned boundary conditions as follows:

- A no-flow boundary is defined on the northern, western, and southern borders, and the aquifer base at -10 m.
- A 10-m-wide river, in the center of the watershed, ranges in stage from 20 to 5 m and is underlain by 5-m-thick sediments with their base at an elevation of 5 m. Rivers are represented as a head-dependent flux boundaries using the MODFLOW-2000 (Harbaugh et al. 2000) river package.
- A 10-m-wide river also bounds the east edge with a stage of 5 m, and 5 m of sediments with their base at 0 m.
- A recharge of 8×10^{-4} m/d is applied uniformly to the top of the model, constituting all the inflow to the system.
- A well pumps 2000 m³/d at $x = 2050$ and $y = 550$.

True heads and flows are generated using a synthetic heterogeneous model with five zones of hydraulic conductivity (K), and a grid of 250×150 cells, each 20×20 m (Figure 1). The model grid used for calibration and prediction consists of 50×30 cells, each 100×100 m (Figure 2). The “true” hydraulic conductivity distribution (Figure 1) includes five zones, with values ranging from 1 to 25 m/d. Vertical hydraulic conductivity of the east-west-oriented riverbed is 0.2 m/d, while that of the north-south riverbed is 0.1 m/d.

Alternative Models

In practice, alternative models should be developed based on careful consideration of the uncertainties associated with understanding of the site hydrology and their representation by the simulation software. For the purpose of illustrating the model evaluation procedure, we generate alternative models by varying the number and

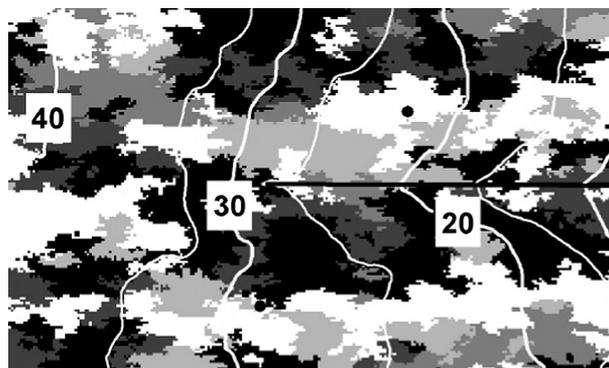


Figure 1. True heterogeneity and head distribution for the synthetic model under hydraulic conditions used to generate calibration data.

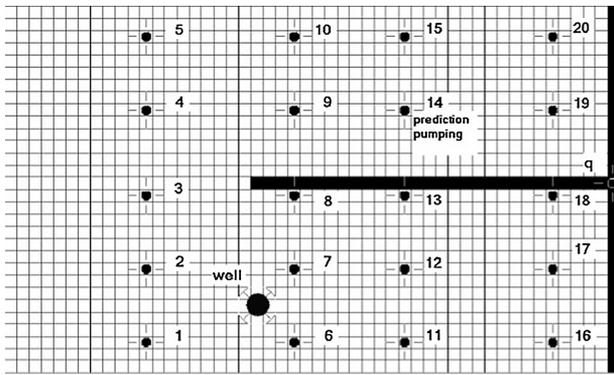


Figure 2. Coarse grid showing rivers (bold lines), observation locations (heads: dots, flux: bracket), and pumping well location.

distribution of hydraulic conductivity zones. Ten sequential indicator simulations (A through J) are generated on the fine grid using GeoStatistical LIBrary (GSLIB) (Deutsch and Journel 1992), the indicator variograms of the synthetic hydrogeologic units, and honoring 144 points of known lithologic type taken from the generating model on a regular grid. Each realization was partitioned into 2 two-zone (e.g., 2A-2J and 2AL-2JL: L indicates that bias is toward low- K material because zone 3 is included with zones 1 and 2 rather than zones 4 and 5), 1 three-zone (3A-3J), 2 four-zone (4A-4J and 4AL-4JL; for L models, zone 3 is included with zone 2 rather than zone 4), and 1 five-zone model (5A-5J); in addition, a homogeneous model was evaluated, resulting in a total of 61 models. In field application the diversity of models will be much greater, including variations of boundary conditions, geologic structure and unit thicknesses, as well as the use of alternative code features to represent features of the ground water system (e.g., in MODFLOW using constant head cells vs. drains, rivers, or streams to simulate communication with surface water).

Calibration data include 20 head observations on a regular grid from the generating model with a hypothesized standard deviation of 0.02-m measurement error and a base flow observation to the central tributary of 6188 m³/d with a standard deviation of 58 m³/d. These standard deviations needed to be increased by a factor of 38 to account for model error and obtain a calculated error variance of 1.0. MODFLOW (Harbaugh et al. 2000; Hill et al. 2000) is used to simulate heads and flows for each model and to estimate a value for K of each zone and the uniform recharge rate; using weights calculated as the inverse of the measurement variance resulted in a dimensionless weighted sum of squared residuals (WSSR). The calibrations require a few seconds on a 3-GHz Pentium 4 PC.

Predictions of flow to both the central tributary and the eastern river and heads at 20 locations, each 200 m up-gradient of the calibration data locations, are made while simulating additional pumping of 3000 m³/d at $x = 3250$ m and $y = 2150$ m. Head distribution in the generating model for the predictive conditions is illustrated in Figure 3.

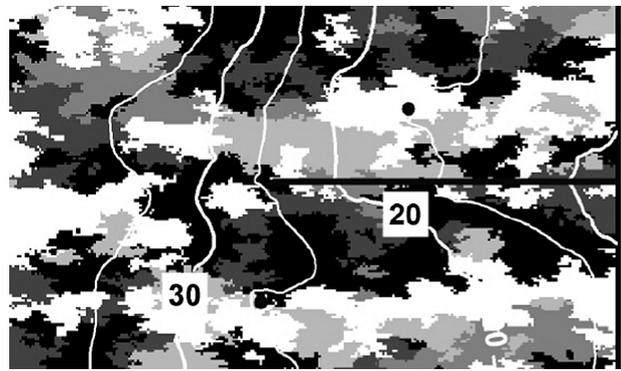


Figure 3. Head distribution in the synthetic model under hydraulic conditions for prediction.

Evaluation Software

J_MMRI is used to evaluate example models. J_MMRI is an early-stage application of the JUPITER (Joint Universal Parameter Identification and Evaluation of Reliability) application programming interface (API), which is currently under development through cooperation of the USGS and U.S. EPA (Poeter et al. 2003). The API provides researchers with open-source program modules and utilities that undertake universal basic tasks required for evaluating sensitivity, assessing data needs, estimating parameters, and evaluating uncertainty, so researchers can focus on developing methods without “reinventing the wheel,” while providing practitioners with public domain software to facilitate the use of the new techniques. J_MMRI collects soft information about each model including (1) model structure: dimensionality, complexity of processes, method of parameter generation/degree of regularization, model representation of features, number/size of model cells/elements, and length/mass/time units; (2) residual distribution: spatial, temporal, and randomness; (3) feasibility of optimal parameter values: absolute and relative; (4) objective function: weighted sum of squares and log likelihood; (5) model selection statistics (e.g., AICc, Bayesian information criterion [BIC], Hannan and Quinn’s criterion [HQ], and Kashyap’s information criterion [KIC]); (6) residual quality: Gaussian character, degree of spatial bias, and similarity to data error; and (7) parameter correlation/certainty. This information is analyzed and organized to facilitate subjective evaluation of the models and provide quantitative model ranking and weighting measures.

Model Ranks

Models were discarded from consideration if the regression did not converge in 20 iterations (two models), or K of a lower-zone number (finer grained material) exceeded the K of a higher-zone number (13 models), leaving 46 of the 61 models for ranking and weighting. It is preferable to include a defensible number of plausible models; however, some models yield unreasonable relative values or cannot be used because they do not converge; thus, the results are not valid for use in further computation. Examinations show that these situations typically occur when the connectivity of hydraulic

conductivity units differ significantly from the true conditions (Poeter and McKenna 1995). For example, if a discontinuous high- K field unit is represented in a model by a continuous unit, then a low- K value may be estimated for the high- K unit in order to compensate for too much continuity. Model selection statistics are given for the best 18 models in Table 1. The number of parameters varied from only three (K for one zone, recharge rate, and σ^2) to seven (K for five zones, recharge rate, and σ^2). σ^2 is counted as a parameter because formally, the likelihood function in the case of normal errors reads as $L(B, \sigma^2|X, g)$ and means “the likelihood of the (unknown) vector of β_s , and, σ^2 , given the data (X) and the model (g).” From the AICc scores, the Δ_i values, and weights, model 4F is the best model, 2J ranks second, and models 5J, 4FL, and 3F have less support, while a number of models have weights of a few percent. The remaining models have relatively little empirical support. Most of the 10 five-zone (seven-parameter) models, are not retained based on unreasonable relative parameter values. Although there are only 21 observations, the more complex models receive high ranks, likely due to the fact that all the geostatistical simulations were well conditioned so the complex models capture the zones well. With less conditioning, simpler models may do a better job of capturing the gross connectivity.

Alternative Model Selection Criteria

We recommend approaches based on K-L information (e.g., AICc) for both model selection and multimodel inference. These methods are based on the concept that models are approximations (i.e., there are no true models of field systems) and select models with more parameters (structure) as the number of observations increase. That

is, in complex systems, smaller effects are identified as the number of observations increase.

There are many other criteria for model selection (McQuarrie and Tsai 1998), and we offer brief comments on some of the alternatives. The BIC (Schwarz 1978), HQ (Hannan and Quinn’s 1979) criterion, and KIC (Kashyap 1982) have been suggested for selection of ground water models (Carrera and Neuman 1986; Neuman 2003; Neuman and Weirenga 2003; Ye et al. 2004). These criteria are similar in form to AICc and are as follows

$$\text{BIC} = n \log(\sigma^2) + k \log(n) \quad (4)$$

$$\text{HQ} = n \log(\sigma^2) + ck \log(\log(n)) \quad \text{where, } c > 2 \quad (5)$$

$$\text{KIC} = n \log(\sigma^2) + k \log\left(\frac{n}{2\pi}\right) + \log|X^T \omega X| \quad (6)$$

where, $|X^T \omega X|$ is the determinant of the Fisher information matrix, X is the sensitivity matrix, X^T is its transpose, and ω is weight matrix.

We do not recommend these procedures as they assume that the true (or quasi-true) model exists in the set of candidate models (Burnham and Anderson 2004), and their goal is to identify this model (as n approaches infinity, probability converges to 1.0 for the true model). These criteria strive for consistent complexity (constant k) regardless of the number of observations. In practice, these criteria can perform similarly to AICc; however, their theoretical underpinnings are philosophically weak. McQuarrie and Tsai (1998) give a readable account of this issue, as do Burnham and Anderson (2002, sections 6.3 and 6.4). Deeper insights are provided in Burnham and Anderson (2004).

Recall that as the number of estimated parameters increases, bias decreases but variance increases (i.e., precision decreases, error bars are larger). The alternative criteria approach the “true model” asymptotically (i.e., as the number of observations increase). However, in most ground water models, the number of observations is small relative to the number of parameters estimated, and these criteria tend to select models that are too simple (i.e., underfitted). Thus, they tend to select for less bias and greater certainty, which threatens to capture a precise but inaccurate answer. We argue that it is preferable to select the model that provides the best approximation to reality for the number of observations available.

A final comment is that AICc and BIC can be derived under either a Bayesian or a frequentist framework. Thus, an argument for or against a criterion should not be based on its Bayesian or frequentist lineage. Rather, one must ask if the true (or quasi-true) model can be expected to be in the set of candidate models in a particular discipline. If so, then criteria such as BIC, HQ, and KIC should be used. In cases where models are merely approximations to complex reality, AICc is preferable (Burnham and Anderson 2002). In addition, AICc has a cross-validation property that is important and stems from its derivation (Stone 1977).

Ranks and model probabilities (weights) for the best 18 models based on AICc are presented in Table 2. BIC

Table 1
Statistics for the 18 Best Models¹ ($n = 21$ in all)

ID**	WSSR	σ^2	k	AICc	Δ_i	w_i
4F	9.40	0.45	5	1.1	0.0	0.2585
2J	13.67	0.65	3	1.5	0.4	0.2155
5J	8.03	0.38	6	2.4	1.3	0.1356
4FL	10.41	0.50	5	3.3	2.1	0.0884
3F	12.82	0.61	4	3.6	2.5	0.0734
3D	13.67	0.65	4	5.0	3.9	0.0374
2H	16.53	0.79	3	5.5	4.3	0.0294
2F	16.68	0.79	3	5.7	4.5	0.0267
5F	9.38	0.45	6	5.7	4.6	0.0262
2A	17.75	0.85	3	7.0	5.8	0.0139
3G	15.25	0.73	4	7.3	6.2	0.0119
2GL	18.11	0.86	3	7.4	6.3	0.0112
2E	18.14	0.86	3	7.4	6.3	0.0111
2C	18.61	0.89	3	8.0	6.8	0.0085
2B	18.71	0.89	3	8.1	6.9	0.0080
4GL	13.17	0.63	5	8.2	7.1	0.0075
2FL	19.17	0.91	3	8.6	7.5	0.0062
4D	13.52	0.64	5	8.8	7.6	0.0057

* The remaining 28 models had essentially zero weight ($<5 \times 10^{-03}$) and are not shown.

** See Alternative Conceptual Models Section for description of model IDs.

Table 2
Weights in Rank Order¹

Model ²	BIC	Model ²	HQ	Model ²	KIC
5J	0.3034	5J	0.4280	5J	0.4549
4F	0.2638	4F	0.2473	4F	0.1884
2J	0.1086	4FL	0.0845	5F	0.0795
4FL	0.0902	5F	0.0828	4FL	0.0717
5F	0.0587	2J	0.0449	2J	0.0695
3F	0.0464	3F	0.0289	3F	0.0276
3D	0.0237	3D	0.0147	3D	0.0165
2H	0.0148	4GL	0.0072	5G	0.0116
2F	0.0134	2H	0.0061	4GL	0.0103
4GL	0.0077	2F	0.0056	2H	0.0074
3G	0.0075	4D	0.0055	2F	0.0067
2A	0.0070	4N	0.0049	5A	0.0066
4D	0.0058	3G	0.0047	3G	0.0065
2GL	0.0057	5B	0.0044	5B	0.0052
2E	0.0056	5G	0.0043	4D	0.0051
4N	0.0052	5A	0.0034	4N	0.0045
2C	0.0043	2A	0.0029	2A	0.0036
2B	0.0040	2GL	0.0023	5D	0.0035

¹ Top 18 ranked models, remaining models had very low weights.
² See "Alternative Conceptual Models" section for description of model labels

and HQ produce results similar to AICc for this particular example, where $n = 21$ and k ranges from only three to seven parameters. The same model is ranked highest by all three measures. The same seven models occupy the top seven ranks (constituting 89%, 93%, and 91% of the weight for BIC, HQ, and AICc, respectively) although in slightly different order. At lower ranks, there is more variation.

Multimodel Inference

The traditional approach to data analysis has been to find the best model, based on some criteria or test result, and make inferences, including predictions and estimates of precision, conditional on this model (as if no other models had been considered). In hindsight, this strategy is poor for a number of reasons. Often, the best model is not overwhelmingly best; perhaps, the weight for the best model is only 0.25 as in Table 1. Thus, there is nonnegligible support for other models. In this case, confidence intervals estimated using the best model are too narrow, and multimodel inference is desirable.

Model Averaging

Model averaging allows estimation of optimal parameter values and predictions from multiple models. Both are calculated in a similar manner; however, we discuss model averaging of predictions first because it is straightforward due to the fact that the same items are predicted using each model, whereas each model may not have the same parameters.

In the example, the best model, 4F, has an AICc weight of only 0.26. This value reflects substantial model uncertainty. If a predicted value differs markedly across the models (i.e., the \hat{y} differs across the models $i = 1,$

$2, \dots, R$), then it is risky to base prediction only on the selected model. An obvious possibility is to compute an estimate of the predicted value, weighting the predictions by the model weights (w_i). This can be done under either a frequentist or Bayesian paradigm. Here, we take the frequentist approach, using K-L information because it is easy to compute and effective in application. If no single model is clearly superior, one should compute model-averaged predictions as

$$\hat{y} = \sum_{i=1}^R w_i \hat{y}_i \quad (7)$$

where \hat{y}_i is the predicted value for each model i , and \hat{y} denotes the model-averaged estimate.

For the estimated regression parameter, $\hat{\beta}_j$, we average over all models where $\hat{\beta}_j$ appears

$$\hat{\beta}_j = \sum_{i=1}^{R'} w'_i \hat{\beta}_{j,i} \quad (8)$$

Thus, the model weights must be recalculated to sum to 1 for the subset of models, R' , that include $\hat{\beta}_j$. When possible, one should use inference based on the subset of models that include $\hat{\beta}_j$ via model averaging because this approach has both practical and philosophical advantages. Where a model-averaged estimator can be used, it appears to improve accuracy and estimates of uncertainty, compared to using $\hat{\beta}_j$ from the selected best model (Burnham and Anderson 2002, section 7.7.5). Parameter averaging is rarely useful for ground water modeling because use of an average parameter value in a particular model construct is not appropriate. However, model-averaged parameter values could provide a range of values for a material type given its multiple representations in alternative models.

Unconditional Variance

Unconditional variance is calculated from multiple models for either parameter values or predictions as shown here for predictions

$$\hat{\text{var}}(\hat{y}) = \left[\sum_{i=1}^R w_i \left[\hat{\text{var}}(\hat{y}_i | \text{model}_i) + (\hat{y}_i - \hat{y})^2 \right]^{0.5} \right]^2 \quad (9)$$

This expression allows for model selection uncertainty to be part of precision because the first term represents the variance, given one calibrated model, and the second term represents the among-model variance, given the set of models. This variance should be used whether the prediction is model averaged or not. The standard deviation is merely the square root of the unconditional variance. Thus, approximate 95% confidence intervals can be expressed using the usual procedure

$$95\% \text{ confidence intervals on } \hat{y} = \hat{y} \pm 2\sqrt{\hat{\text{var}}(\hat{y})} \quad (10)$$

If a $\hat{\beta}_j$ is to be averaged across models where it appears, the number of models (R') and the recalculated model weights (w'_i) must be used in expressions

equivalent to Equations 9 and 10, with $\hat{\beta}$ replacing \hat{y} , and β_j replacing \hat{y}_i .

Extended Example

None of the models considered for the example problem is clearly “the best” as indicated by the AICc weight of 0.26 for the model with the highest rank. The evidence ratio for the best and second best models indicates that the best model is only 1.2 times more likely than the second model, given the evidence, indicating a lack of strong support for the best model. A model needs to have a weight greater than ~0.95 before considering it as the best model and bypassing the multimodel averaging process.

Predictive Quality of the Best Model and the Multimodel Average

Desirable predictive models are those with small weighted mean square error (WMSE) between their predictions and those of the generating model. The prediction locations are 200 m upgradient (left) of the calibration locations, which are shown in Figure 2. Weighting by the inverse of measurement variance, using the same weights for heads and flows as used in the calibration to account for differences in magnitude and units of measurements, predictive WMSE is calculated. WMSE is the sum of the mean weighted squared differences between the 20 heads and 2 flows predicted by the alternative models and the true heads and flows simulated by the generating model with the additional pumping. The correlation between AICc and KIC model ranks and that WMSE for the 46 retained models illustrates the best fit to calibration data does not assure the most accurate predictions at all locations in the model (Figure 4). This is also illustrated by the relationship of the WSSR for the calibration and the WMSE for predictions (Figure 4). It has been noted that ground water models with the best fit to calibration data will not necessarily produce the most accurate predictions (Yeh and Yoon 1981; Rushton et al. 1982).

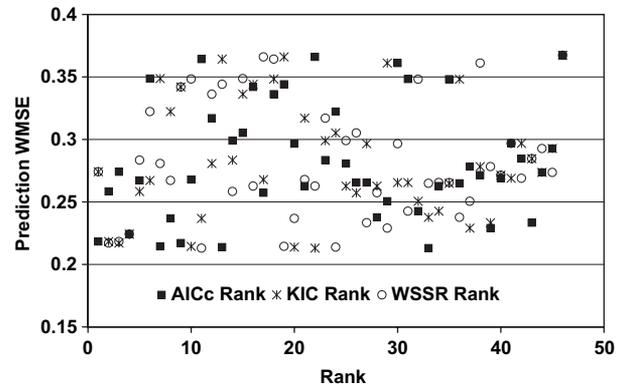


Figure 4. Relationship between AICc, WSSR, and KIC ranks and predictive WMSE.

Rigorous experimental comparison of the alternative model ranking criteria requires evaluation of many different systems and numerous realizations of observation sets that is beyond the scope of this paper. Such an exercise would only reveal empirical value of the alternative methods because their theoretical underpinning is not well founded, as we know it is impossible to include the true model of a ground water system in the set of models.

Predictions at most locations are fairly accurate and readily captured by the linear confidence intervals of most of the alternative models. Model-averaged head predictions and their Scheffe confidence intervals are presented in Figure 5 for each individual model at locations 7, 8, and 10 (located 200 m upgradient of the calibration points with the same ID in Figure 2). At location 7, nearly all models underestimate head, and confidence intervals of the top four models do not capture the truth. Model averaging (Equations 9 and 10) increases the confidence intervals and captures the truth (i.e., the value predicted by the generating model) (Figure 5a). Although large numbers of simulations would be needed to make a rigorous statistical evaluation, the practical similarity of the approaches is illustrated by noting the following: of the 22

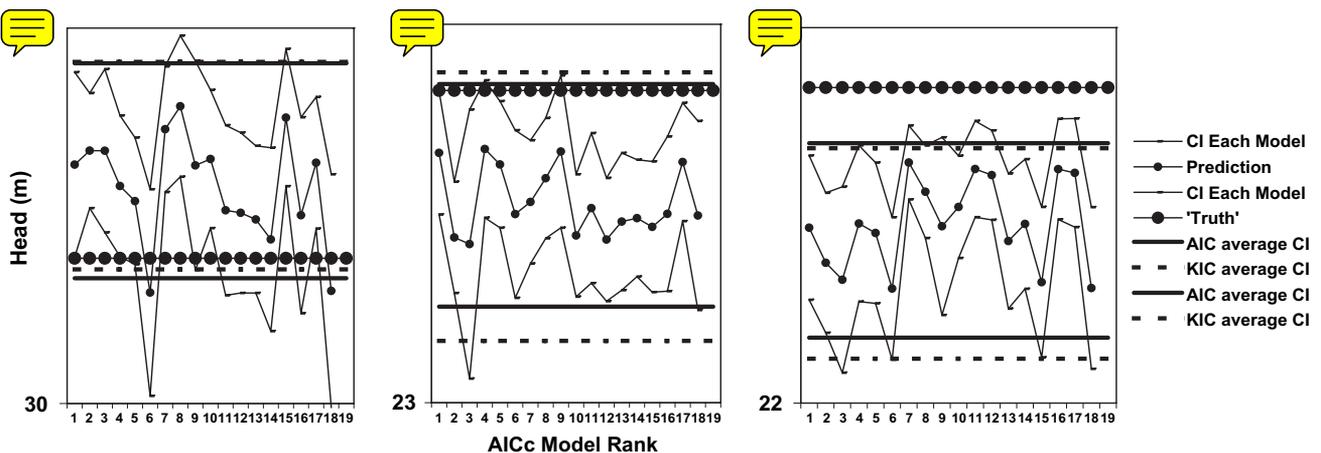


Figure 5. Predictions for location numbers 7, 8, and 10 (see Figure 2 for locations) are examples of locations where the models tend to (a) underestimate, (b) overestimate, and (c) inaccurately estimate the head, respectively, under the new pumping conditions. The predicted value and linear individual confidence intervals (CI) are shown for each model, as are model-averaged confidence intervals based on AICc and KIC.

predictions made for this example, 21 were captured by the model-averaged intervals. The predicted head at location 1 (Figure 5b) tends to be overestimated, and some of the best ranked models barely capture the truth in the confidence intervals based on individual model variance, but model averaging clearly captures the truth. Head at location 8 (Figure 5c) is not predicted successfully by any of the models and this is so consistent that it is not captured by model averaging. Although large numbers of simulations would be needed to make a rigorous statistical evaluations, the practical similarity of the approaches is illustrated by noting the following: of the 22 predictions made for this example, 12 were captured by the model-averaged intervals.

In this example, predictions generated by the alternative models vary considerably in some locations and not in others. This is illustrated in Figure 6 where the difference between the high and low head of all 46 models is displayed as a function of location. It is also indicated by noting that model averaging increases the confidence intervals indicated by the best model on 1 of the 22 predictions by less than 25% and another by 166%, with an average increase of 72% and a median of 64%. This variability serves to increase model-averaged variance through the second term of Equation 9, which is carried forward to confidence intervals in Equation 10. Field applications are likely to exhibit more striking variation in models including differences in geometry and boundary conditions, hence more significant shift of prediction and broadening of confidence intervals as a result of model averaging.

Summary and Conclusions

Given our uncertainty of site conditions, hydrologists should routinely consider several, well-thought-out models to maintain an open mind about the system. Generally, inferences should stem from multiple plausible models (multimodel inference) because it yields more robust predictions and a more “honest,” realistic measure of precision. Modelers should be keenly aware of the fact that even multimodel inference, which provides greater consideration for uncertainty, is vulnerable to yielding poor predictions if fundamentally important processes are not included in the model, predictive locations and/or conditions differ substantially from those of the calibration, or the prediction horizon is large relative to the calibration time frame as discussed by Bredehoeft (2003).

Multimodel ranking and inference approaches based on K-L information, such as the AICc measure presented here, are simple to compute, easy to interpret, and provide a rigorous foundation for model-based inference. 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

+1.5	+4.2	+2.4	+1.7
+1.5	+2.4	+1.8	+3.0
+1.4	+1.7	+0.8	+2.0
+2.0	+1.3	+1.0	+1.3
+2.5	+1.1	+0.8	+1.2

Figure 6. Total head drop across the model is 35 m, while the difference between the highest and lowest predicted head of the 46 models at a given location ranges from 0.8 to 4.2 m.

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 application, it is unreasonable to assume that one 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.

Acknowledgments

The final manuscript was improved based on suggestions of Neil Blanford, Steffen Mehl, and Jim Yeh. Their reviews are greatly appreciated. U.S. EPA and USGS provided funding for this work, but responsibility for its content lies with the authors.

References

- Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control AC* 19, 716–723.
- Akaike, H. 1973. Information theory as an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, ed. B.N. Petrov, 267–281. Budapest, Hungary: Akademiai Kiado.
- Anderson, D.R. 2003. Multi-model inference based on Kullback-Leibler information. In *Proceedings of MODFLOW and More 2003: Understanding Through Modeling, IGWMC*, 366–370.
- Bredehoeft, J.D. 2003. From models to performance assessment: The conceptual problem. *Ground Water* 41, no. 5: 571–577.
- Burnham, K.P., and D.R. Anderson. 2004. Multi-model inference: Understanding AIC and BIC model selection. *Sociological Methods and Research* 33, no. 2: 261–304.
- Burnham, K.P., and D.R. Anderson. 2002. *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. New York: Springer-Verlag.
- Carrera, J., and S.P. Neuman. 1986. Estimation of aquifer parameters under transient and steady state conditions: 1. Maximum likelihood method incorporating prior information. *Water Resources Research* 22, no. 2: 199–210.
- Chamberlin, T.C. 1890. The method of multiple working hypotheses. *Science* 15: 93–98. Reprinted 1965, *Science* 148: 754–759.

- deLeeuw, J. 1992. Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle. In *Breakthroughs in Statistics*, vol. 1, ed. S. Kotz and N.L. Johnson, 599–609. London, U.K.: Springer-Verlag.
- Deutsch, C., and A. Journel. 1992. *GSLIB: Geostatistical Software Library and User's Guide*. New York: Oxford University Press.
- Hannan, E.J., and B.G. Quinn. 1979. The determination of the order of an autoregression. *Journal of the Royal Statistical Society Series B* 41, no. 1: 190–195.
- Harbaugh, A.W., E.R. Banta, M.C. Hill, and M.G. McDonald. 2000. MODFLOW-2000 the U.S. Geological Survey modular ground water model—User guide to modularization concepts and the ground water flow process. USGS Open-File Report 00–92. Reston, Virginia: USGS.
- Hill, M.C., E.R. Banta, A.W. Harbaugh, and E.R. Anderman. 2000. MODFLOW-2000, the U.S. Geological Survey modular ground water model—User guide to the observation, sensitivity, and parameter-estimation processes and three post-processing programs. USGS Open-File Report 00–184. Reston, Virginia: USGS.
- Hurvich, C.M., and C.-L. Tsai. 1994. Autoregressive model selection in small samples using a bias-corrected version of AIC. In *Engineering and Scientific Applications*, vol. 3, ed. Bozdogan, H. 137–157. Proceedings of the First U.S./Japan Conference on the Frontiers of Statistical Modeling: An Informational Approach. Dordrecht, Netherlands: Kluwer Academic Publishers.
- Hurvich, C.M., and C.-L. Tsai. 1989. Regression and time series model selection in small samples. *Biometrika* 76, no. 2: 297–307.
- Kashyap, R.L. 1982. Optimal choice of AR and MA parts in autoregressive moving average models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 4, no. 2: 99–104.
- Kullback, S., and R.A. Leibler. 1951. On information and sufficiency. *Annals of Mathematical Statistics* 22, 79–86.
- Linhart, H., and W. Zucchini. 1986. *Model Selection*. New York: John Wiley and Sons.
- McQuarrie, A.D.R., and C.-L. Tsai. 1998. *Regression and Time Series Model Selection*. Singapore: World Scientific Publishing Company.
- Neuman, S.P. 2003. Maximum likelihood Bayesian averaging of uncertain model predictions. *Stochastic Environmental Research and Risk Assessment* 17, no. 5: 291–305.
- 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). Washington, D.C.: U.S. Nuclear Regulatory Commission.
- Parzen, E., K. Tanabe, and G. Kitagawa, eds. 1998. *Selected Papers of Hirotugu Akaike*. New York: Springer-Verlag.
- Poeter, E., M. Hill, J. Doherty, J.E. Banta, and J. Babendreier. 2003. JUPITER Project—Joint Universal Parameter Identification and Evaluation of Reliability, Fall AGU Meeting. Abstract H12G-03. *Eos* 84, no. 46: F608–F609.
- Poeter, E.P., and M.C. Hill. 1997. Inverse methods: A necessary next step in ground water modeling. *Ground Water* 35, no. 2: 250–260.
- Poeter, E.P., and S.A. McKenna. 1995. Reducing uncertainty associated with groundwater flow and transport predictions. *Ground Water* 33, no. 6: 899–904.
- Rushton, K.R., E.J. Smith, and L.M. Thomlinson. 1982. An improved understanding of flow in a limestone aquifer using field evidence and a mathematical model. *Journal of the Institute for Water Engineers and Scientists*. 36, no. 5: 369–383.
- Sakamoto, Y., M. Ishiguro, and G. Kitagawa. 1986. *Akaike Information Criterion Statistics*. Tokyo, Japan: KTK Scientific Publishers.
- Schwarz, G. 1978. Estimating the dimension of a model. *Annals of Statistics* 6, no. 2: 461–464.
- Stone, M. 1977. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, Series B* 39, no. 1: 44–47.
- Sugiura, N. 1978. Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics, Theory and Methods* A7: 13–26.
- Ye, M., S.P. Neuman, and P.D. Meyer. 2004. Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. *Water Resources Research* 40, no. 5: W05113, 1–19.
- Yeh, W.W.-G., and Y.S. Yoon. 1981. Aquifer parameter identification with optimum dimension in parameterization. *Water Resources Research* 17, no. 3: 664–672.