Evaluation of Plausibility of Alternative Groundwater Models Using Different Kinds of Observations

Ming Ye[1], Liying Wang[1], Karl F. Pohlmann[2]
[1]*Florida State University, mye@fsu.edu, Tallahassee, FL, USA*
[2]*Desert Research Institute, Karl.Pohlmann@dri.edu, Las Vegas, NV, USA*

ABSTRACT

There has been a growing trend in groundwater modeling to use information criteria and/or model selection criteria to evaluate model plausibility. However, the criteria cannot be used as the sole means for model evaluation. The most direct and fundamental evidence of model plausibility should be based on analysis of model fit to observations. Using multiple kinds of observations is particularly useful, since different kinds of observations contain different information about the system of interest. We present in this paper a case study in which model plausibility based on observations is opposite to that based on commonly used information criteria such as AIC, BIC, and KIC. The modeled area is the Death Valley Regional Flow System (DVRFS) located in Nevada and California. While a total of twenty-five alternative models were developed in a previous study, this study is focused on the six most plausible cases, a combination of two recharge and three geological models. In addition to field observations of hydraulic head and discharge and estimates of constant-head boundary flow that were used in previous studies, an estimate of interbasin flow was used as a new constraint. The best fit to the field observations and the interbasin flow estimate was achieved by using the Morris method for sensitivity analysis and Monte Carlo method for calibration. Based on the parameter set corresponding to the best fit, AIC, BIC, and KIC identified the best model, which has the best overall goodness-of-fit to the observations, dominated mainly by the goodness-of-fit to head observations. This model, however, simulates the wrong flow direction along several segments of the constant-head boundary in order to maintain mass balance. It suggests that evaluation of model plausibility should not be based on overall model fit but on that of individual kinds of observations.

INTRODUCTION

Hydrologic analyses are commonly based on a single conceptual/mathematical model, yet hydrologic environments are open and complex, rendering them prone to multiple interpretations and conceptualizations. This is true regardless of the quantity and quality of available hydrologic information and data. With recognition of this, there has been a growing trend in groundwater modeling to conduct multimodel analysis in which predictive analysis is not based on a single model but on multiple models that are plausible given available data and information (Neuman, 2003; Ye et al., 2004; Poeter and Anderson, 2005; Refsgaard et al., 2006). A special issue published in Stochastic Environmental Research and Risk Assessment may help define the state of the art in quantification of model uncertainty (Ye et al., 2010a). A common practice for evaluating plausibility of alternative models is to first calibrate the models and then calculate the following widely used information criteria (AIC and AICc) and/or Bayesian model selection criteria (BIC and KIC). For a given criterion, a model with the smallest value is considered the most plausible one. In multimodel analysis, the criteria are further used to calculate model averaging weight or model probability for conducting model averaging.

While it has been shown that the criteria are capable of selecting the best model and improving prediction in multimodel analysis, the criteria should be used with caution in real-world applications and should not be used as the sole means of evaluating model plausibility. Real-world situations are complex and may not conform fully to assumptions behind the derivations of the criteria. For example, Ye et al. (2008, 2010b) showed that, in the derivation of BIC and KIC, several assumptions are made but they may not be fully satisfied in reality. These include the validity of disregarding higher-order terms in the derivation of KIC, which would render the likelihood function more complex and non-Gaussian; ignoring cross correlations between the models and their prior parameter estimates; and/or misrepresenting prior data

and parameter statistics. These introduce sufficient ambiguity into the analysis to justify relying on multiple model selection criteria as has become the norm in recent practice.

In comparison with the model selection criteria, model fit to observations contains more information about model plausibility and should be carefully examined before using the model selection criteria in multimodel analysis. In particular, different kinds of observations contain different information about the system of interest, which enables us to evaluate model plausibility from the hydrogeologic angle. This is demonstrated in this paper through groundwater modeling at the Death Valley Regional Flow System (DVRFS), located in southwest Nevada and the Death Valley area of California (Figure 1a). Groundwater flow modeling in the DVRFS is of national importance, because predicting radionuclide transport at the Department of Energy's Nevada National Security Site (NNSS) is critical for protecting human health and the environment. With respect to multimodel analysis, Pohlmann et al. (2007) and Ye et al. (2010c) developed a total of twenty-five groundwater flow models, as a combination of five recharge and five geological models. These models were calibrated within the modeling framework of Belcher et al. (2004), and model plausibility was evaluated using AIC, BIC, and KIC based on the calibration results. This study is an extension of the previous study with focus on how to use model fit for evaluating model plausibility.

ALTERNATIVE RECHARGE AND GEOLOGICAL MODELS

Among the twenty-five groundwater flow models developed in Pohlmann et al. (2007) and Ye et al. (2010c), only six models were considered in this study, based on previous model calibration and expert judgment. The six groundwater models are related to two recharge models and three hydrostratigraphic framework models (HFMs). Figures 1b and 1c plot estimate of net infiltration from the net infiltration with runon-runoff (NIM1 or R2) and estimate of recharge from the chloride mass-balance with alluvial and elevation masks (CMB2 or R5) (the model abbreviations are adopted from Ye et al., 2010c). Although the estimates of the two models have similar patterns, their values are dramatically different. Figure 2 compares the three HFMs: DVRFS model (G1), UGTA base model (G2), and UGTA CP-Thrust model (G3), where UGTA stands for Underground Test Area. The second model is an update of the first model at the northern Yucca Flat area. As illustrated in the north-south cross-section in Figures 2a and 2b, the two models differ in both the number of hydrostratigraphic units and their subsurface configuration. The third alternative model incorporates a different interpretation of the configuration of hydrostratigraphic units with respect to the CP thrust fault; this alternative model was developed based on the UGTA base model

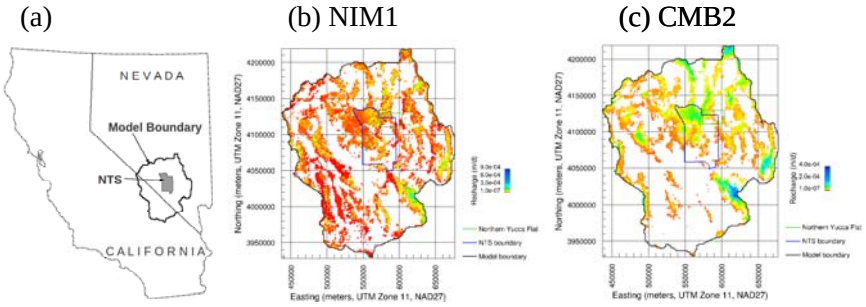

Figure 1. (a) Boundaries of the Death Valley Regional Flow System and the Nevada National Security Site, and (b) and (c) recharge rate estimates (m/d) of the two recharge models.

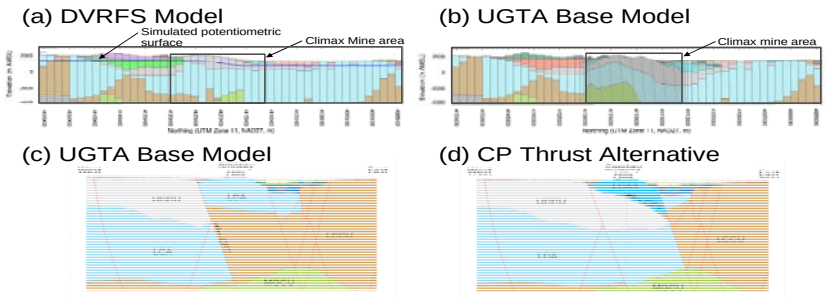

Figure 2. Two-dimensional illustration of difference between (a and b) the DVRFS and UGTA base models, and (c and d) the UGTA base model and the CP Thrust alternative.

to address uncertainty regarding the particular features of hydrostratigraphy that might be important to groundwater flow and contaminant transport in Yucca Flat.

MODEL CALIBRATION AND RESULTS

Forward and inverse modeling was conducted within the DVRFS model framework developed using MODFLOW 2000 (Belcher et al., 2004). However, the original transient model was converted to steady state, representing conditions prior to groundwater development. Simulation times for running the steady state models are dramatically reduced, which makes possible more comprehensive inverse modeling and MC simulations. The steady-state model is developed from the transient model by removing and revising components related to transient-state simulations. For example, the parameters related to specific storage and observations of head-changes are removed in the steady-state models.

The field observations used to calibrate the steady-state models include 700 observations of hydraulic heads, 45 observations of discharges, and 15 constant-head boundary flows. In addition, because simulation of interbasin flow is critical to this modeling, an estimate of interbasin flow, Q_y, into northern Yucca Flat was also incorporated in the calibration. However, because of the use of MODFLOW 2000, the estimate was used only as a constraint on model simulations, not as an explicit calibration target. The inter-basin flow estimate is subject to large uncertainty, ranging from 1,180 m^3/d (Winograd and Thordarson, 1975) to more than 100,000 m^3/d (Pohlmann et al., 2007). This uncertainty is due to uncertainty in the methods used to estimate the inter-basin flow, (2) uncertainty in model parameters (especially when the number of calibrated parameters is large), and (3) uncertainty between different conceptual models. The estimate of 25,000 m^3/d is considered reasonable (IT Corporation, 1996) and is used in this study.

Model calibration was conducted in two steps to match the field observations and estimate of inter-basin flow. The six models were first calibrated against the three kinds of field observations in the manner of Pohlmann et al. (2007) using the statistics of measurement errors given in Belcher et al. (2004). Although the calibration results, measured by SSWR (sum of squared weighted residuals), are reasonable, the interbasin flow is overestimated. This is due to sensitivity of Q_y to model parameter estimates. A centered parameter study shows that Q_y can be reduced at potential cost of SSWR. In order to select a set of model parameters with simulated Q_y honoring the estimates, a Monte Carlo (MC) simulation was conducted. To avoid MC simulation of large parameter space (about 50 to 60 parameters were calibrated), a screening-level sensitivity analysis using the Morris method was conducted to select parameters to which Q_y simulations are sensitive. A total of six parameters were selected for models related to G1, but nine (the same) for models G2 and G3.

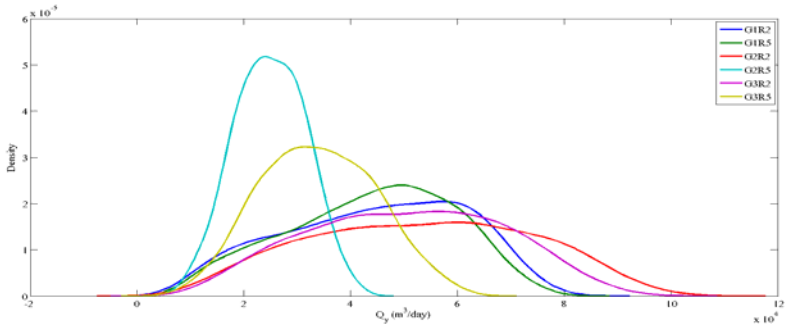

Figure 3. Probability density functions of the Q_y estimates for the six models.

Model	G1R2	G1R5	G2R2	G2R5	G3R2	G3R5
Inter-basin flow (m^3/d)	24,444	25,743	22,672	19,173	22,988	21,747
SSWR (total)	12,491	12,380	11,560	10,678	11,455	11,170
SSWR(head only)	11,970	11,602	10,986	9,859	10,864	10,379
SSWR(discharge only)	451.86	651.72	532.51	611.91	553.76	650.62
SSWR(constant-head boundary flow only)	68.96	125.82	42.12	205.78	37.19	140.51

Table 1. Optimum inter-basin flow and SSWR values for each model.

Subsequently, 2,000 Monte Carlo simulations were conducted for the critical parameters by assuming that the parameters follow uniform distribution with assumed ranges varying around the parameter estimates. Figure 3 plots the probability density of Q_y for the six models. It shows that both parametric uncertainty and model uncertainties are significant in the Q_y estimate. Since this study was focused on model uncertainty, only one realization of each model was selected based on the criteria that (1) the Q_y estimate should be close to 25,000 m^3/d and (2) the SSWR should be the smallest after the first criterion is satisfied. The Q_y estimate, total SSWR, and SSWR of each kind of observations are listed in Table 1 for the selected realization of the six models. The model with the best goodness-of-fit is G2R5, whose SSWR of head is the smallest but the SSWR of constant-head boundary flow is the largest among the six models. The latter SSWR plays critical roles in evaluation of model plausibility as discussed below.

MODEL PLAUSIBILITY ANALYSIS

Based on the selected realization for each model, AIC, BIC, and KIC and their corresponding model probabilities were evaluated and listed in Table 2. According to the model posterior probability listed in Table 2, model G2R5 is considered as the most plausible model with almost 100% probability by all the three criteria. The reason is that this model has the smallest SSWR; penalty terms in the model selection criteria thus have negligible effect on model selection. Without conducting more hydrologic analysis and relying solely on the statistics, one would choose model G2R5 for predictive analysis.

Model	AIC		BIC		KIC	
	Value	$P(M_k\|D)$	Value	$P(M_k\|D)$	Value	$P(M_k\|D)$
G1R2	2229.58	0.00	2465.88	0.00	2322.10	0.00
G1R5	2218.79	0.00	2445.83	0.00	2280.03	0.00
G2R2	2184.71	0.00	2453.44	0.00	2270.27	0.00
G2R5	2106.39	1.00	2333.43	1.00	2198.87	0.97
G3R2	2177.78	0.00	2446.51	0.00	2265.86	0.00
G3R5	2140.63	0.00	2367.66	0.00	2205.90	0.03

Table 2. AIC, BIC, and KIC and corresponding model probability.

However, comparing simulated constant-head boundary flow with the observed indicates that model G2R5 is not a physically reasonable model. Table 3 shows that, for four segments of the constant-head boundary, the simulated boundary flows of G2R5 have opposite direction to those observed. This explains why the SSWR of constant-head flow of model G2R5 is the largest. Since the SSWR of constant-head flow is small (because of small weights used for the flow observation), it has small contribution to the overall SSWR dominated by the SSWR of head. However, the opposite flow direction is a direct evidence that the model is physically unreasonable. This is also true for all models associated with recharge model R5 and model G1R2.

Observation name	Observation	G1R2	G1R5	G2R2	G2R5	G3R2	G3R5
C_LASV0303	-3633	-2989	-8152	-2165	-7319	-2081	-7561
C_SHPR0401	-4410	-2203	-2552	-2281	-3709	-2001	-3569
C_SHPR0402	-15305	-24207	-33660	-22419	-42755	-20134	-43702
C_SHPR0403	-4959	-6107	-9687	-5113	-11459	-4665	-12323
C_SHPR0404	5927	4637	9111	4221	14225	3632	15839
C_PAHR0501	1827	2480	3074	2299	5440	2047	5765
C_PAHR0502	-2346	-1657	-2255	-14	-999	-341	-930
C_PAHR0505	-2521	-10207	-12398	-7245	-13393	-7149	-12370
C_GRDN0603	2334	1873	-5176	645	-9064	653	-7239
C_STNC0700	12476	62155	28548	72061	-17039	64116	3382
C_CLAY0800	667	-1019	-1796	1381	-4275	1639	-2691
C_EURS0900	15100	4495	24499	938	16413	491	24818
C_PANA1100	15000	16095	16560	23289	6906	23771	10514
C_OWLS1203	1682	3593	4383	4474	2941	4559	3339
C_SILU0100	500	-2060	-3363	82	-2400	343	-2456

Table 3. Observed and simulated values of constant-head boundary flow.

The reason that the models related to recharge model R5 are physically unreasonable is that the recharge estimate of R5 is too excessive to maintain mass balance with the observed discharge and boundary flow. For model G2R5, its recharge estimate is 361,075m^3/d, observed discharge and boundary flow are -328,546m^3/d and 22,339m^3/d, respectively. Before the calibration, the mass balance error is 54,868m^3/d. During the calibration, outflow is forced on the inflow boundary to maintain mass balance.

SUMMARY

In this paper we present a case study of DVRFS modeling to demonstrate that the information criteria (AIC) and model selection criteria (BIC and KIC) cannot be used as the sole means for evaluating plausibility of alternative models. AIC, BIC, and KIC unanimously select the same best model, because the model's overall SSWR is the smallest due to the best model fit to head observations. However, the selected model is not physically reasonable because it simulates incorrect flow directions along certain segments of the constant-head boundary in order to maintain mass balance. Evaluation of model plausibility should not be based on overall model fit but on model fit of individual kinds of observations, because different kinds of observations contain different information on the system of interest.

ACKNOWLEDGEMENTS

This work was funded by the U.S. Department of Energy under Contract Number DE-AC52-06NA26383 and was also supported in part by NSF-EAR grant 0911074.

REFERENCES

Belcher, W.R. (ed), 2004. Death Valley regional ground-water flow system, Nevada and California – Hydrogeologic framework and transient ground-water flow model, U.S. Geological Survey Scientific Investigation Report 2004-5205.

IT Corporation, 1996. Groundwater Flow Model Documentation Package, Volume VI in Underground Test Area Subproject, Phase I, Data Analysis Task, ITLV/10972--181.

Neuman, S.P., 2003. Maximum likelihood Bayesian averaging of alternative conceptual-mathematical models, Stochastic Environmental Research and Risk Assessment, 17(5), 291-305, DOI: 10.1007/s00477-003-0151-7.

Poeter, E. and Anderson, D.R., 2005. Multimodel ranking and inference in groundwater modeling, Ground Water, 43 (4), 597-605.

Pohlmann, K., M. Ye, D. Reeves, M. Zavarin, D. Decker, and J. Chapman, 2007. Modeling of Groundwater Flow and Radionuclide Transport at the Climax Mine sub-CAU, Nevada Test Site, DRI Publication 45226, DOE/NV/26383-06, Nevada Site Office, National Nuclear Security Administration, U.S. Department of Energy, Las Vegas, NV.

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

Winograd, I.J. and Thordarson, W., 1975. Hydrogeologic and Hydrochemical Framework, South-Central Great Basin, Nevada-California, with Special Reference to the Nevada Test Site. U.S. Geological Survey Professional Paper 712-C.

Ye, M., Neuman, S.P. and Meyer, P.D., 2004. Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff, Water Resources Research, 40 (5), W05113, doi:10.1029/2003WR002557.

Ye, M., Meyer, P.D., and Neuman, S.P., 2008. On model selection criteria of multimodel analysis, Water Resources Research, Water Resources Research, 44, W03428, doi:10.1029/2008WR006803.

Ye, M., Meyer, P.D., Lin, Y.-F., and Neuman, S.P., 2010a. Quantification of model uncertainty in environmental modeling, Stoch. Environ. Res. Risk Assess., doi:10.1007/s00477-010-0377-0.

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

Ye, M., Pohlmann, K.F., Chapman, J.B., Pohll, G.M., Reeves, D.M., 2010. A model-averaging method for assessing groundwater conceptual model uncertainty. Ground Water, doi:10.1111/j.1745-6584.2009.00633.x.