ground water

A Model-Averaging Method for Assessing Groundwater Conceptual Model Uncertainty

by Ming Ye[1], Karl F. Pohlmann[2], Jenny B. Chapman[2], Greg M. Pohll[3], and Donald M. Reeves[3]

Abstract

This study evaluates alternative groundwater models with different recharge and geologic components at the northern Yucca Flat area of the Death Valley Regional Flow System (DVRFS), USA. Recharge over the DVRFS has been estimated using five methods, and five geological interpretations are available at the northern Yucca Flat area. Combining the recharge and geological components together with additional modeling components that represent other hydrogeological conditions yields a total of 25 groundwater flow models. As all the models are plausible given available data and information, evaluating model uncertainty becomes inevitable. On the other hand, hydraulic parameters (e.g., hydraulic conductivity) are uncertain in each model, giving rise to parametric uncertainty. Propagation of the uncertainty in the models and model parameters through groundwater modeling causes predictive uncertainty in model predictions (e.g., hydraulic head and flow). Parametric uncertainty within each model is assessed using Monte Carlo simulation, and model uncertainty is evaluated using the model averaging method. Two model-averaging techniques (on the basis of information criteria and GLUE) are discussed. This study shows that contribution of model uncertainty to predictive uncertainty is significantly larger than that of parametric uncertainty. For the recharge and geological components, uncertainty in the geological interpretations has more significant effect on model predictions than uncertainty in the recharge estimates. In addition, weighted residuals vary more for the different geological models than for different recharge models. Most of the calibrated observations are not important for discriminating between the alternative models, because their weighted residuals vary only slightly from one model to another.

Introduction

Groundwater modeling is commonly based on a single conceptual model. Yet groundwater environments are open and complex, rendering them prone to multiple interpretations and conceptualizations. This is particularly true for regional-scale modeling, in which parameter measurements and field observations are sparse relative to large modeling domains. It is not uncommon for new data to invalidate prevailing conceptual models, and it is difficult to select a single appropriate conceptual model (Bredehoeft 2003, 2005). There is a growing tendency among groundwater modelers to postulate alternative models for a site. Neuman and Wierenga (2003) elucidated various situations in which multiple models are needed. The most often encountered situations are different models of alternative descriptions of groundwater processes and interpretations of hydrogeological data (Sun and Yeh 1985; Carrera and Neuman 1986; Samper and Neuman 1989; Harrar et al. 2003; Tsai et al. 2003; Ye et al. 2004, 2008a; Poeter and Anderson 2005; Foglia et al. 2007; Troldborg et al. 2007; Rojas et al. 2008, 2009; Tsai and Li 2008). Evaluation of the alternative models becomes inevitable when the models are all acceptable (to various extents) given available knowledge and data. Ignoring conceptual

[1]Corresponding author: Department of Scientific Computing, Florida State University, Tallahassee, FL 32306; (850) 644-4587; fax (850) 644-0098; mye@fsu.edu

[2]Division of Hydrological Sciences, Desert Research Institute, Las Vegas, NV 89119.

[3]Division of Hydrological Sciences, Desert Research Institute, Reno, NV 89512.

Received December 2008, accepted August 2009.

doi: 10.1111/j.1745-6584.2009.00633.x

model uncertainty may result in biased predictions and/or underestimation of predictive uncertainty. Model averaging has received increasing attention for assessing model uncertainty (Neuman 2003; Ye et al. 2004; Hojberg and Refsgaard 2005; Poeter and Anderson 2005; Refsgaard et al. 2006, 2007), and general purpose computer software that implements model averaging has been developed (Poeter and Hill 2007). However, the majority of studies on model averaging are limited to synthetic cases, in which model complexity can be controlled but is significantly simpler than that of real world conditions. This study incorporates real world conditions to enhance our understanding of model uncertainty and the model averaging methods currently in use for assessing model uncertainty.

The main purpose of this study is to investigate two sources of model uncertainty for real world groundwater modeling under complicated hydrogeological conditions. One source of uncertainty arises from the difficulty of choosing, from multiple alternatives, an appropriate model for estimating groundwater recharge (or net infiltration). The rationale for using multiple recharge models is that it may increase reliability of recharge estimates given recharge model uncertainty (Scanlon et al. 2002). The other source of uncertainty arises from uncertainty in geological models due to different interpretations of geological and geophysical data, a well-known contributor to model uncertainty. In this study, recharge estimated using different methods and different geological interpretations is incorporated in a groundwater modeling framework. This leads to alternative groundwater models, and they are evaluated simultaneously by using expert judgment (through expert elicitation) and on-site observations (through model calibration). The relative importance of the two sources of model uncertainty to groundwater flow modeling is discussed; overall model uncertainty is assessed using a model averaging method as discussed below.

Propagation of model uncertainty through groundwater modeling gives rise to predictive uncertainty, as different models lead to different model predictions (e.g., hydraulic head and flow). The predictive uncertainty is also attributed to propagation of parametric uncertainty. For each of the models, hydraulic parameters are uncertain due to spatial variability of the parameters and to paucity of parameter measurements and field observations used for model calibration. In this study, a large number of hydraulic parameters are calibrated, and corresponding parametric uncertainty in the calibrated hydraulic conductivity is assessed using Monte Carlo methods. As a result, for each model, predictive uncertainty is reflected by multiple realizations of model predictions. When alternative models are considered, predictive uncertainty is quantified by aggregating predictive uncertainty of each model using the model averaging method. In other words, the results of model averaging quantify both model uncertainty and parametric uncertainty. This attempt of comprehensively assessing predictive uncertainty for a complex, real world groundwater model has not been reported previously in the literature.

The study site is the northern Yucca Flat area of the Nevada Test Site, located within the Death Valley Regional Flow System (DVRFS), USA (Figure 1). Three underground nuclear tests were conducted between 1962 and 1966 in the Climax mine granite stock immediately north of Yucca Flat. Groundwater flow and contaminant transport modeling is now under way to estimate radionuclide flux from the Climax stock to the downgradient Yucca Flat in support of corrective action investigations by the U.S. Department of Energy (DOE). This paper focuses only on the flow modeling in northern Yucca Flat; the associated transport modeling is described in Pohlmann et al. (2007) and Reeves et al. (2009). Because there are only 59 hydraulic head observations in northern Yucca Flat (Figure 1) and boundary conditions in the area are largely unknown, the groundwater flow modeling is conducted within the DVRFS modeling framework (Belcher et al. 2004) implemented using MODFLOW-2000 (Harbaugh et al. 2000; Hill et al. 2000). Hereinafter, the DVRFS modeling framework is referred to as the DVRFS model. The DVRFS model has been developed over the last decade on the basis of regional characterization of hydraulic, geological, and hydrogeological conditions of the DVRFS. In addition, this MODFLOW model

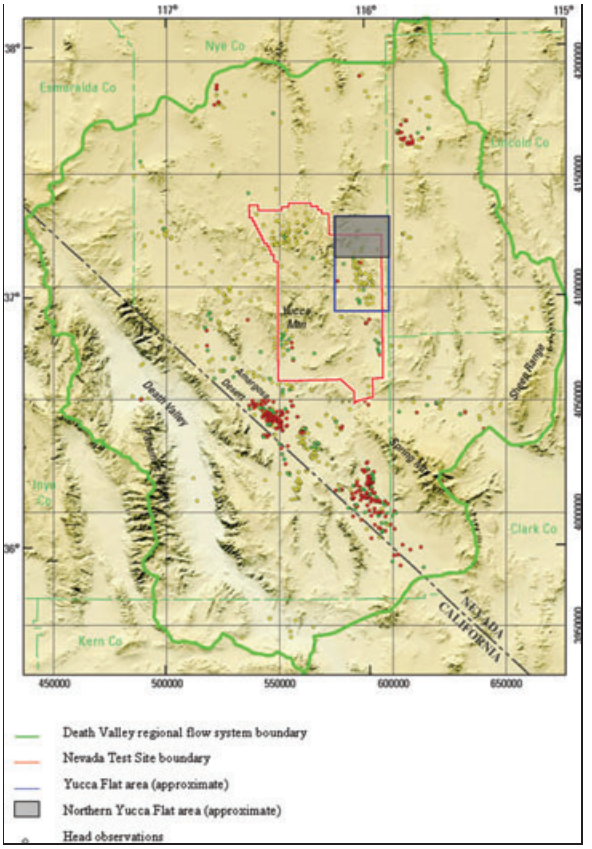

Figure 1. Map showing boundaries of the Death Valley Regional Flow System (DVRFS), the Nevada Test Site, and the northern Yucca Flat (area of detailed geological models). The figure is modified from Belcher et al. (2004).

has been calibrated against 4963 observations of head, head-change, and flow (locations of the head observations are shown in Figure 1). Using the DVRFS modeling framework, rather than developing a separate flow model for northern Yucca Flat, is expected to better constrain the flow system in northern Yucca Flat.

Complicated hydrogeologic conditions in the DVRFS lead to significant uncertainties in its recharge and geological components. To date, recharge at the DVRFS has been estimated using five methods. In addition, five geological interpretations (rigorously speaking, hydrostratigraphic frameworks) have been developed by Belcher et al. (2004) and Bechtel Nevada (2006) for northern Yucca Flat. Several geological interpretations are necessary to reflect nonunique understandings of geological and geophysical data resulting from the complexity of the model area. Although a single combination of the recharge and geological components was used in the DVRFS model, an expert elicitation suggested that there is no justification for selecting a single recharge/geological model and discarding others (Ye et al. 2008a). The five recharge and five geological components are thus all used for modeling groundwater flow in northern Yucca Flat. Although there is no question that other kinds of model uncertainty exist in the DVRFS, discussion of all sources of model uncertainty is beyond the scope of this study. Similarly, this study focuses on parametric uncertainty of the hydraulic conductivity and ignores other sources of parametric uncertainty.

The goal of this study is to incorporate the uncertainty in the recharge and geological components into groundwater flow predictions. As illustrated in the figure of the Supporting Information, by replacing the recharge and geological components of the DVRFS model with the alternatives, a total of 25 alternative groundwater flow models are produced (as the result of combinations of the five recharge and five geological models). Plausibility of the alternative models is evaluated by calibrating the models against the same observation data and by estimating probabilities of the individual models using the method discussed below.

The conventional model averaging method is used to assess the model uncertainty. If Δ is the quantity of interest predicted by a set of K alternative models, then its distribution conditioned on a dataset $\boldsymbol{D}$ is (Hoeting et al. 1999):

$$p(\Delta|\boldsymbol{D}) = \sum_{k=1}^{K} p(\Delta|M_k, \boldsymbol{D}) p(M_k|\boldsymbol{D}) \tag{1}$$

where $p(\Delta|M_k, \boldsymbol{D})$ is the predictive probability of Δ for model M_k, and $p(M_k|\boldsymbol{D})$ is the posterior probability of M_k. One way of estimating the posterior model probability, the averaging weight, is to use Bayes' theorem:

$$p(M_k|\boldsymbol{D}) = \frac{p(\boldsymbol{D}|M_k)p(M_k)}{\sum_{l=1}^{K} p(\boldsymbol{D}|M_l)p(M_l)} \tag{2}$$

where $p(\boldsymbol{D}|M_k)$ is the likelihood of model M_k (a measure of consistency between model predictions and site observations $\boldsymbol{D}$) and $p(M_k)$ is prior probability of M_k. In this study, instead of using a noninformative equal prior [$p(M_k) = 1/K$], informative prior model probabilities are obtained from expert elicitations (Ye et al. 2008a). This is expected to improve predictive performance of the model averaging (Ye et al. 2005). In general, the first two moments of Δ are used to quantify the uncertainty. For modelM_k, parametric uncertainty is quantified by the mean, $E[\Delta|\boldsymbol{D}, M_k]$, and variance, $\mathrm{Var}[\Delta|\boldsymbol{D}, M_k]$, which can be obtained using either Monte Carlo simulation or stochastic methods. The posterior mean and variance

$$E[\Delta|\boldsymbol{D}] = \sum_{k=1}^{K} E[\Delta|\boldsymbol{D}, M_k] p(M_k|\boldsymbol{D}) \tag{3}$$

$$\begin{aligned}\mathrm{Var}[\Delta|\boldsymbol{D}] = &\sum_{k=1}^{K} \mathrm{Var}[\Delta|\boldsymbol{D}, M_k] p(M_k|\boldsymbol{D}) \\ &+ \sum_{k=1}^{K} (E[\Delta|\boldsymbol{D}, M_k] - E[\Delta|\boldsymbol{D}])^2 p(M_k|\boldsymbol{D})\end{aligned} \tag{4}$$

quantify both the parametric and model uncertainty. The first and second terms on the right-hand side of Equation 4 are the within- and between-model variance, respectively.

In groundwater modeling, a common practice for estimating the model likelihood function, $p(\boldsymbol{D}|M_k)$, is to approximate it using model selection (or information) criteria: AIC (Akaike 1974), AICc (Hurvich and Tsai 1989), BIC (Schwarz 1978), or KIC (Kashyap 1982). Definitions of each of the criteria are given as the supporting information. A general approximation that includes all criteria can be expressed as:

$$p(\boldsymbol{D}|M_k) = \exp\left(-\frac{1}{2}\mathrm{IC}_k\right) \tag{5}$$

where IC_k is any of the four criteria of model M_k. When the least-square method is used for model calibration, the first term of the model selection criteria becomes (Poeter and Anderson 2005; Ye et al. 2008b):

$$\begin{aligned}-\ln[L(M_k|\boldsymbol{D})] &= N \ln \hat{\sigma}^2_{k,ML} = N \ln(\boldsymbol{e}_k^T \boldsymbol{\omega} \boldsymbol{e}_k / N) \\ &= N \ln(\mathrm{SSWR}_k / N)\end{aligned} \tag{6}$$

where N is number of calibration data $\boldsymbol{D}$, $\mathrm{SSWR}_k = \boldsymbol{e}_k^T \boldsymbol{\omega} \boldsymbol{e}_k$ is sum of squared weighted residual of model M_k, $\boldsymbol{e}_k$ is the vector of residuals (difference between observations and simulations of modelM_k), and ω is the weight matrix due to measurement error of the observations. Another way of approximating the likelihood function is given by the generalized likelihood uncertainty estimation (GLUE) method (see a review by Beven [2006]). Although multiple expressions are given in Beven and Binley (1992) and Beven and Freer (2001), this study uses the likelihood measure

$$p(\boldsymbol{D}|M_k) = (\boldsymbol{e}_k^T \boldsymbol{\omega} \boldsymbol{e}_k)^{-E} \tag{7}$$

where E is a parameter chosen by the user. When $E = 0$, all models will have the same likelihood; and when $E \to \infty$, the best model with the smallest SSWR will have the posterior model probability of 1. Rojas et al. (2008) used several other likelihood definitions in the model averaging context.

In the remaining part of this paper, the alternative recharge estimation methods and geological interpretations are briefly described, followed by discussion of the model calibration results. Before giving the concluding remarks at the end of this paper, calculation of the posterior probabilities of the groundwater models and effect of the calibration data on the calculation are elaborated.

Alternative Recharge and Geological Components

The alternative recharge estimation methods and geological interpretations are briefly described in this section; further details are presented in Pohlmann et al. (2007) and their original publications.

Five Recharge Estimation Methods

The five recharge models of the DVRFS are:

(1) MME (R1): modified Maxey-Eakin method. This method is based on the empirical Maxey-Eakin method that estimates groundwater recharge as a function of precipitation estimates for selected zones of elevation (Maxey and Eakin 1949), and is updated using new methodologies and datasets (Epstein 2004) and an expanded area of coverage to include the entire Death Valley region.

(2 and 3) NIM1 (R2) and NIM2 (R3): net infiltration methods with/without runon-runoff. The two methods employ a distributed-parameter watershed model for estimating temporal and spatial distribution of net infiltration and potential recharge (Hevesi et al. 2003). The difference between the two methods is that R2 incorporates a surface water runon-runoff component whereas R3 does not.

(4 and 5) CMB1 (R4) and CMB2 (R5): chloride mass balance methods with alluvial mask (R4) and both alluvial and elevation masks (R5) (Russell and Minor 2002). The two methods estimate groundwater recharge on the basis of the elevation-dependent chloride mass balances within hydrologic input and output components of individual hydrologic basins. In R4, recharge in areas covered by alluvium is eliminated (the alluvial mask). In addition to the alluvial mask, R5 further eliminates recharge in areas below 1237 m elevation (the elevation mask).

The five methods are based on three different recharge estimate techniques: an empirical approach (MME), an approach based on unsaturated-water studies (NIM1 and NIM2), and an approach based on saturated-water studies (CMB1 and CMB2). The models reflect different levels of complexity. The empirical MME method is the simplest one; the NIM methods are the most complicated because they consider the processes controlling net infiltration and potential recharge. Figure 2 illustrates the recharge rate estimates of the five methods. The MME gives the highest recharge estimate, and the NIM models give the lowest. Because of the runon-runoff component considered in NIM1, the recharge estimate of NIM1 is higher than that of NIM2, although spatial patterns of the recharge estimate are similar in the two methods. Because of the extra elevation mask considered in CMB2, the recharge estimate of CMB2 is lower than that of CMB1; for the same reason, spatial patterns of the recharge estimate are different in the two models (less recharge is estimated in southern Nevada in CMB2).

Five Geological Interpretations

The five geological interpretations in northern Yucca Flat are:

1. USGS (G1): USGS interpretation (Belcher et al. 2004)
2. BAS (G2): UGTA base interpretation (Bechtel Nevada 2006)

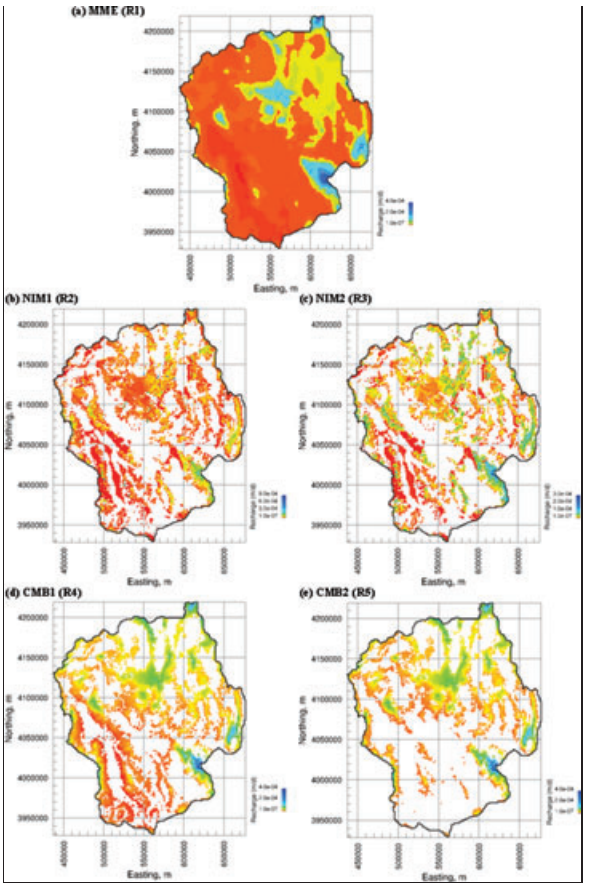

Figure 2. Recharge rate estimates (m/d) of the five recharge models (R1 to R5): (a) modified Maxey-Eakin method (MME), (b) net infiltration method with runon-runoff (NIM1), (c) net infiltration method without runon-runoff (NIM2), (d) chloride mass balance method with alluvial mask (CMB1), and (e) chloride mass balance method with alluvial and elevation masks (CMB2).

3. CPT (G3): UGTA CP thrust alternative (Bechtel Nevada 2006)
4. HB (G4): UGTA hydrologic barrier alternative (Bechtel Nevada 2006)
5. CPT+HB (G5): combination of the UGTA CP thrust and hydrologic barrier alternatives (Bechtel Nevada 2006)

Figure 3 illustrates the major differences between these interpretations in two-dimensional cross sections. The USGS interpretation (G1), developed by the U.S. Geological Survey, represents the configuration of hydrogeologic units in the entire DVRFS. The UGTA base interpretation (G2), developed by Bechtel Nevada (2006) as part of the underground test area (UGTA) program of the Nevada test site, focuses on more local-scale geological and geophysical data and information than the DRVFS. As illustrated in the north-south cross section in Figures 3a and 3b, G1 and G2 differ in both the number of hydrostratigraphic units and their subsurface configuration. The UGTA alternative interpretations (G3 to G5) are developed in response to nonunique interpretations of particular features that may be important to groundwater flow and contaminant transport in northern Yucca Flat. Figures 3c to 3f show the difference between the G2 and its two alternatives. G3 (Figure 3d) incorporates a different interpretation of the configuration of hydrostratigraphic units with respect to the CP thrust fault. The lower carbonate-rock aquifer (LCA) and upper clastic-rock confining unit (UCCU) are extended eastward to replace the lower carbonate-rock confining unit (LCCU) and the LCA located above the LCCU, respectively. As shown in the next section, the flow field corresponding to G3 is dramatically different from that of G2, due to the configuration of the CP thrust. G4 (Figure 3f) postulates a barrier to groundwater flow on the east side of the Climax stock for the purpose of limiting southward groundwater flow from northern basins into northern Yucca Flat to be consistent with observed hydraulic gradients. G5, the combination of G3 and G4, is not shown in Figure 3. As shown below, the geological components are more important than the recharge components for controlling groundwater flow, and the uncertainty of the former dominates over that of the latter in the predictive uncertainty.

Model Calibration Results

The 25 groundwater models (resulting from incorporating the five recharge and five geological components) are calibrated in a similar manner to the calibration of the DVRFS model described in Belcher et al. (2004). Unlike Belcher et al. (2004) who calibrated parameters distributed throughout the entire DVRFS, only parameters in northern Yucca Flat and its vicinity are calibrated in this study. The calibrated parameter values (listed in Pohlmann et al. [2007]) are different (significantly for certain parameters) for different groundwater models. The values assigned to other parameters and variables needed

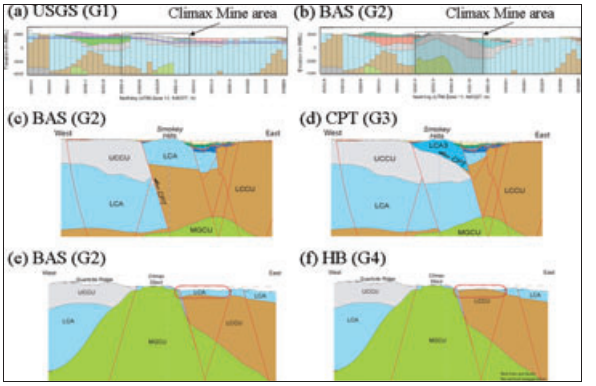

Figure 3. Two-dimensional vertical cross sections illustrating differences between (a and b) the USGS (G1) and UGTA base (G2) interpretations, (c and d) the UGTA base interpretation (G2) and the CP thrust alternative (G3), and (e and f) the UGTA base interpretation (G2) and the hydrologic barrier alternative (G4). Figures (c) and (e) represent two cross sections of G2. Coordinates in (a) and (b) are Northing (meters, UTM Zone 11, NAD27) and elevation (meters).

for the calibration (e.g., convergence criteria and weighting matrix, ω) are adopted from the DVRFS model. Therefore, the model calibration of this study can be viewed as a further calibration of the DVRFS model. Given the new information and data included in the geological models and the local optimization method used in MODFLOW-2000, as discussed in the supporting information, the new calibration improves the calibration and reduces the SSWR of the DVRFS model. Nevertheless, fixing the calibrated parameters from the DVRFS model is a limitation in this study, and it would be more accurate to calibrate all parameters to which simulations of the alternative models are sensitive. For further discussion of the calibrated parameters and their effect on model simulations, readers are referred to Pohlmann et al. (2007).

To compare the residuals of the 25 groundwater models in northern Yucca Flat, all residuals of the 59 observations are plotted together in Figure 4 in a manner that allows for comparison of residuals for each of the five recharge models within a single geological model. Weights were applied to the two types of observations (head and head changes) to render them dimensionless for comparison. The figure shows that, for any geological interpretations, most of the residuals are nearly the same for all the recharge estimation methods, except for five observations of head and two observations of head-change. For the two head-change observations (with indices 58 and 59 in Figure 4), their residuals for most of the 25 models are nearly the same, except for the combinations of G3R3, G1R2, and G1R3. The residuals of the five head observations (with indices 43, 44, 48, 49, and 50 in Figure 4) are significantly different for all of the 25 models, suggesting that head observations are more sensitive than head-change observations to the alternative models. This is also seen in Table S2 of the supporting information, which shows that the SSWR of the head observations varied more dramatically than that of the

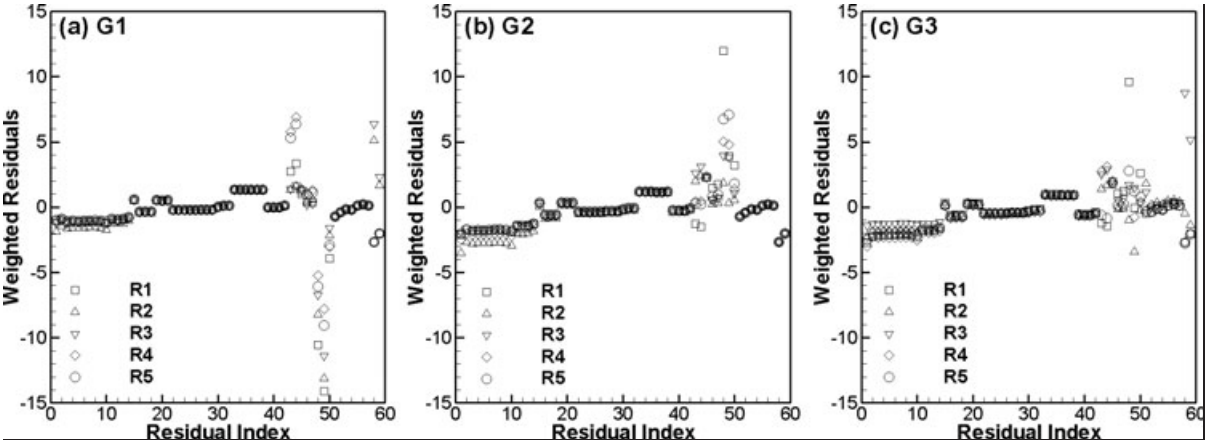

Figure 4. Weighted residuals of the 50 observations of head (index from 1 to 50) and 9 observations of head change (index from 51 to 59) for all of the 25 models. The residuals for different recharge models are plotted together for a single geological model.

head-change observations for different alternative models. Therefore, Figure 4 suggests that the head observations play a more important role of discriminating the models than the head-change observations.

Calculation of Posterior Model Probabilities

Calculation of the posterior model probability using Equation 2 requires the prior model probability and the model likelihood function. The prior probabilities of the 25 models are elicited from two independent expert panels for the recharge and geological models, respectively. Details of the expert elicitation are given in Ye et al. (2008a). The prior model probabilities, plotted in Figure 5, reflect the panelists' beliefs regarding relative plausibility of each model, considering their consistency with available data and knowledge. For the recharge estimation methods, the NIM1 method (R2) has the largest prior probability, indicating that the experts had the most confidence in this method. The confidence is based on the method's comprehensive incorporation of the processes controlling net infiltration and potential recharge.

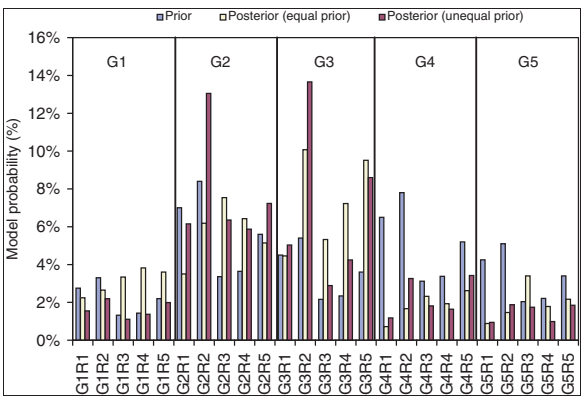

Figure 5. Prior and posterior probabilities of the 25 models. The posterior model probabilities correspond to equal and unequal prior model probabilities.

Not surprisingly, for the geological interpretations, the UGTA base interpretation (G2 or BAS) has the largest prior probability. The CPT alternative (G3) is considered less plausible than the HB alternative (G4) based on drill-hole and geophysical data. As a result of combining the geological and recharge components, model G2R2 has the largest prior probability of 8.4%, about twice as large as the average prior probability of 4% (recall that summation of the prior probabilities of the 25 models is 100%). Figure 7 shows that the prior model probabilities are generally evenly distributed among the 25 models, indicating that there is no justification for selecting a single model (and discarding others) based on the prior judgment from one modeler or a group of modelers.

Table 1 lists the model selection criteria calculated based on the model calibration results using the method of Ye et al. (2008b) as well as the posterior model probabilities estimated using Equations 2 and 5. The table lists only the results for models associated with geological interpretations G2 and G3, because the probabilities of all other models are calculated to be near zero. For all the model selection criteria, only two of the 25 models receive non-negligible probabilities. AIC, AICc, and BIC select models G3R5 and G3R2. The probability of G3R5 is larger than that of G3R2, although the SSWR of G3R2 is smaller than that of G3R5. The reason is that G3R2 has two more calibrated parameters than G3R5. KIC selects models G3R2 and G2R3. Model G2R3 is selected because its log determinant of the Fisher information matrix is significantly smaller than that of model G3R5, despite the fact that the SSWR and number of calibrated parameters of G2R3 are larger than those of model G3R5. From a theoretical point view, calculation of the model selection criteria using observations only in northern Yucca Flat may not be appropriate for this study because model calibration is conducted for the entire DVRFS. In this case, it is unknown whether the theoretical basis of the model selection criteria is sound, especially for the penalty terms (e.g., the $2N_k$ term of AIC and $N_k \ln N$ term of BIC) of the criteria.

Table 1
The AIC, AICc, BIC, and KIC Values and Corresponding Model Probabilities for Models Associated with the Two Geological Models (G2 and G3) and All Recharge Models

	MME (R1)	NIM1 (R2)	NIM2 (R3)	CMB1 (R4)	CMB2 (R5)
UGTA base model (BAS or G2)					
AIC	98.33	68.74	57.10	62.51	75.63
$p(M_k\|\boldsymbol{D})$ (%)	0	0	0	0	0
AICc	101.21	73.32	61.68	65.39	78.51
$p(M_k\|\boldsymbol{D})$ (%)	0	0	0	0	0
BIC	112.87	87.44	75.80	77.05	90.17
$p(M_k\|\boldsymbol{D})$ (%)	0	0	0	0	0
ln|F|	3.53	−13.95	−21.13	−6.33	−5.61
KIC	64.99	12.51	−4.54	23.56	35.84
$p(M_k\|\boldsymbol{D})$ (%)	0	0.02	33.83	0	0
UGTA CP thrust alternative model (CPT or G3)					
AIC	84.11	39.98	77.59	55.56	39.37
$p(M_k\|\boldsymbol{D})$ (%)	0	42.36	0	0.02	57.61
AICc	86.99	44.56	82.18	58.44	42.25
$p(M_k\|\boldsymbol{D})$ (%)	0	23.88	0	0.02	76.10
BIC	98.65	58.68	96.29	70.10	53.91
$p(M_k\|\boldsymbol{D})$ (%)	0	8.43	0	0.03	91.54
ln|F|	3.76	−7.01	−19.70	−4.49	−3.94
KIC	52.69	−4.92	14.26	19.27	5.56
$p(M_k\|\boldsymbol{D})$ (%)	0	65.92	0	0	0.23

Note: Probabilities of other models are zero, and thus are not listed.

Although it is legitimate to conduct model averaging using the model probabilities given by the four model selection criteria, using only 2 of the 25 models for model averaging does not appear to adequately support the main purpose of this study: incorporating model uncertainty in simulations of flow and radionuclide transport from the Climax stock area to downgradient Yucca Flat. Figure 6 plots the simulated flow rate of the six models that are considered the most plausible using the GLUE methodology discussed below. The figure shows apparent differences in flow patterns simulated by geological interpretations G2 and G3. The large flow rate at the bottom of Figures 6a and 6c for G3 (UGTA CP Thrust alternative) corresponds to the LCA unit extending eastward in G3 (Figures 3c and 3d). One would expect to average the flow rates simulated by the two geological models, whereas AIC, AICc, and BIC select only G3R2 and G3R5 associated with geological model G3. Although the two geological models (G3R2 and G2R3) are selected by KIC, KIC still discards all other models. Figure 7 shows that the flow rates simulated by the 25 calibrated models are significantly different, and the simulated values of models G3R2, G3R5, or G2R3 are higher than the average. Therefore, using only two model results for averaging may result in underestimation of predictive uncertainty and bias the predictions, the two problems that model averaging intends to avoid.

Given that the calibration metrics (SSWR) are similar, the GLUE method is used to calculate the posterior probabilities of the 25 models (through Equations 2 and 7). The results plotted in Figure 5 show two sets of posterior probabilities for equal ($1/25 = 4\%$) and unequal (obtained from the expert elicitations) prior probabilities. The GLUE posterior model probabilities are more evenly distributed than those of IC-based model selection criteria; all models receive probabilities within a generally comparable range, the largest and smallest values being 14% and 1%, respectively. Therefore, using the GLUE model probability for averaging the simulated flow avoids the problems above caused by using the IC-based model probabilities. However, it is worth mentioning that the GLUE model probabilities are exclusively based on the goodness-of-fit (the SSWR) without taking into account model complexity.

Figure 5 also shows the effect of prior probability on posterior model probability. Because recharge estimation method R2 (NIM1) and geological interpretation G2 (UGTA base model) are ranked the highest from the expert elicitation, the posterior probabilities associated with R2 and G2 are higher when they are calculated using the unequal prior model probabilities than when using the equal priors. For example, the posterior probability of model G2R2 increases from 6% to 13% when the unequal prior probability is used. It is also important to note that the magnitudes of prior and posterior model probabilities are not always proportional. For instance, the elicitation judged R1 (MME) to be the second best recharge model, whereas the posterior probabilities of models associated with R1 are very low. This is also the case for the geological models, because the best calibration resulted

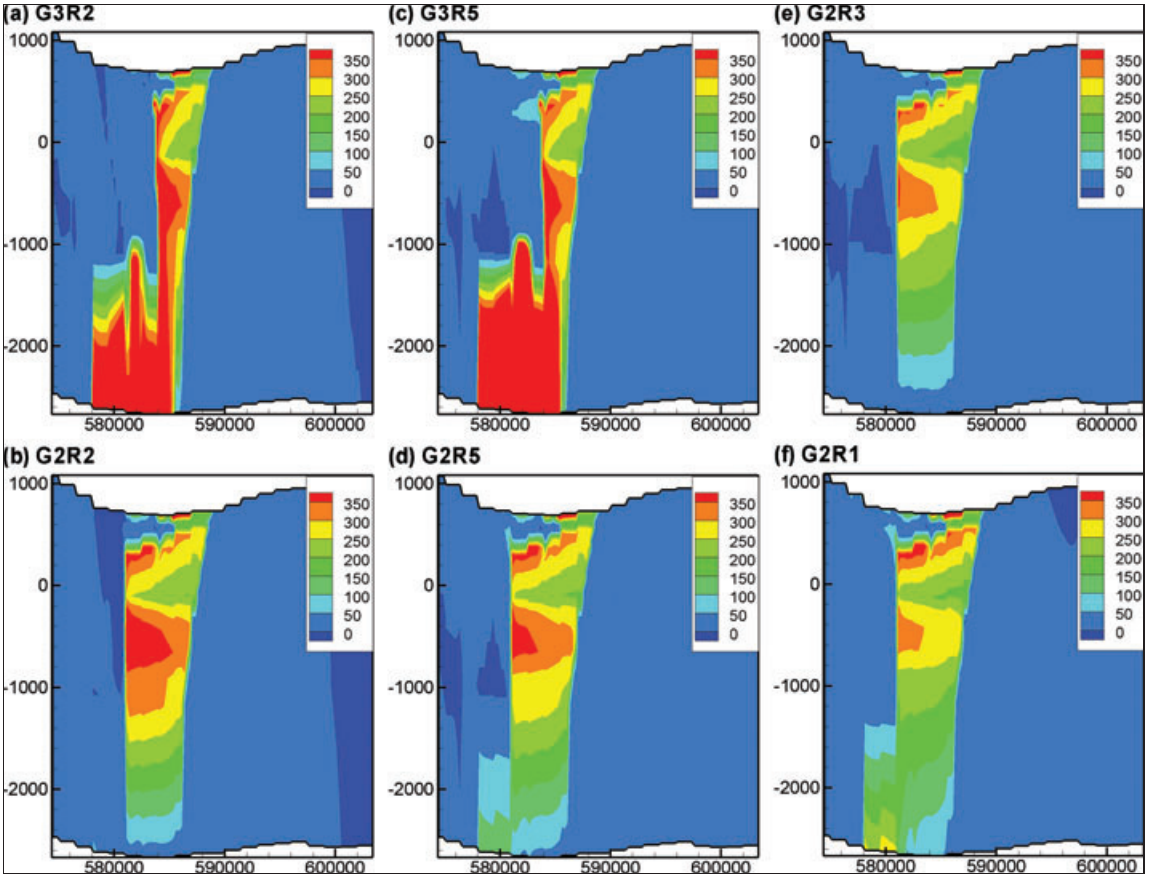

Figure 6. Cell-by-cell flow crossing the front face of MODFLOW cells (i.e., Q_y, m^3/day) predicted at the southern boundary ($x-z$ cross section) of northern Yucca Flat by models (a) G3R2, (b) G2R2, (c) G3R5, (d) G2R5, (e) G2R3, and (f) G2R1. G2 and G3 are the UGTA base model and its CPT alternative. R1, R2, R3, and R5 are MME, NIM1, NIM2, and CMB2 recharge models, respectively. The x- and z-coordinates are Easting (meters, UTM Zone 11, NAD 27) and elevation (meters).

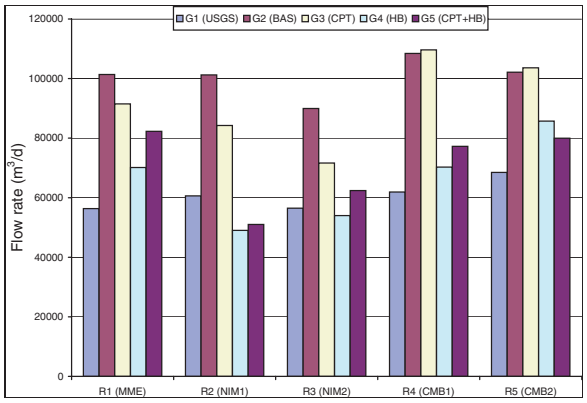

Figure 7. Simulated flow rates from the northern Yucca Flat area into Yucca Flat. Each cluster represents the flow rate of the five geological models for a single recharge model.

from G3 (the UGTA CP thrust alternative) not from G2 (the UGTA base interpretation) that was rated highest in the elicitation. Quantitative analysis regarding the effect of prior model probability on model predictions will be studied in the future.

Effect of Calibration Data on Model Discrimination

Figure 8 is a scatterplot of the weighted residuals of head and head-change observations in northern Yucca Flat given by the six most plausible models selected by the GLUE model probabilities (model probability gradually decreases from G3R2 of Figure 8a to G2R1 of Figure 8f). In these plots, the symbol size is proportional to the magnitude of the residuals. Head contours at the top layer of the domain are also plotted. The figure shows that, in the trough area (in blue) where hydraulic head is low, the magnitude of the residuals is similar. This is also observed in Figure 4 where these residuals vary only slightly for different models. Because model probabilities based on these residuals are similar, calibration against these data is not critical to discriminate between the alternative models. In contrast, observations in the area north of the low-head area are critical, because their values change significantly for the different models. For example, the residual at the location highlighted by the blue circle in Figure 8a increases gradually from Figures 8a to 8f. Correspondingly, the model probabilities gradually decrease from model G3R2 in Figure 8a to G2R2 in Figure 8f.

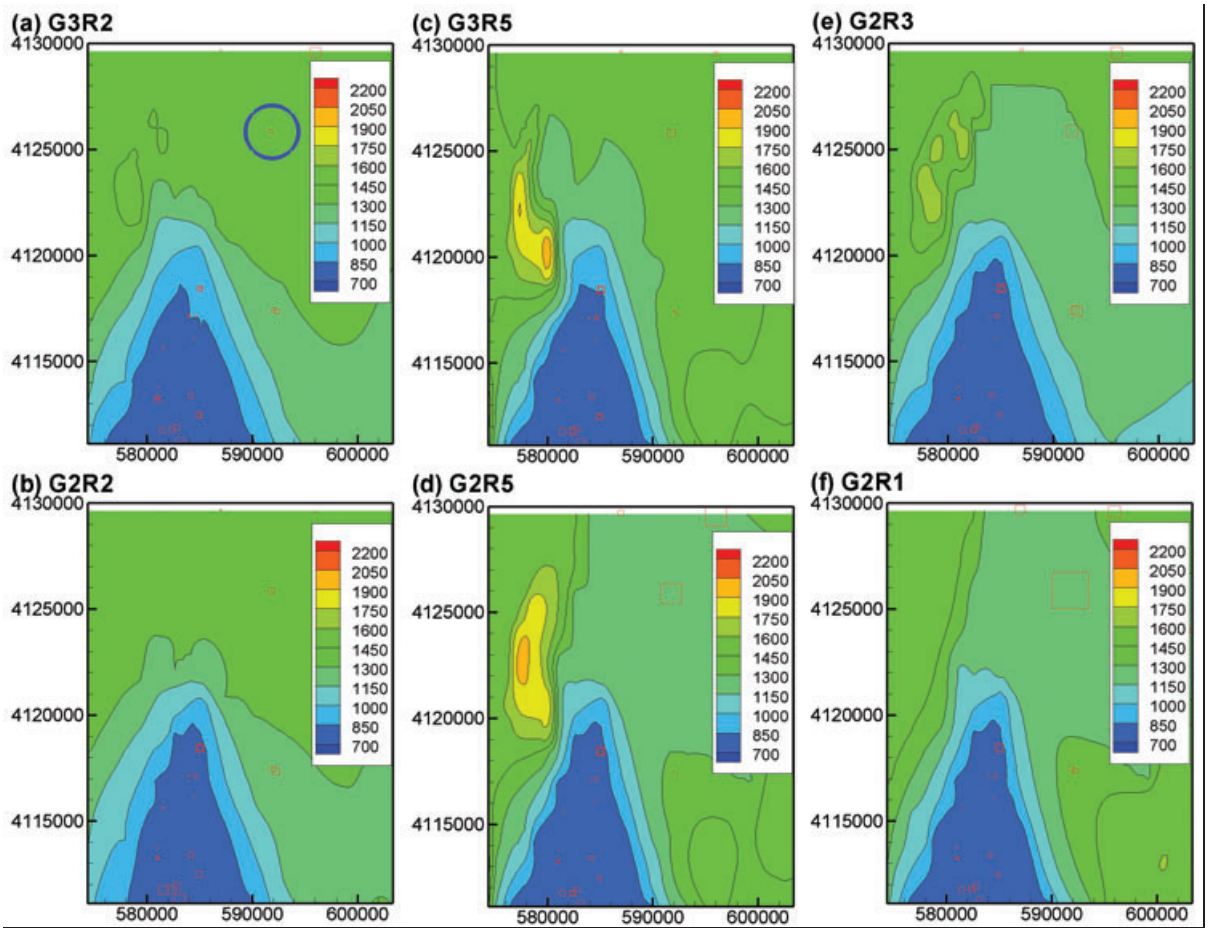

Figure 8. Scatterplots of residuals of heads and head-changes and contours of hydraulic head at the top layer in the northern Yucca Flat area simulated by models (a) G3R2, (b) G2R2, (c) G3R5, (d) G2R5, (e) G2R3, and (f) G2R1. G2 and G3 are the UGTA base model and its CPT alternative. R1, R2, R3, and R5 are MME, NIM1, NIM2, and CMB2 recharge models, respectively. The x- and z-coordinates are Easting and Northing (meters, UTM Zone 11, NAD 27). Size of the scatters is scaled with magnitude of the residuals.

This finding is important for guiding data collection for further discrimination between alternative models and for reducing model uncertainty. For this study, collecting more data in the trough area (in blue) would not help discriminate the models, because they would not reveal more information about differences between alternative models. Instead, more data upgradient of the trough area should be collected, because observations in this area vary significantly for different models. Future studies that focus on the development of statistical measures that quantify the amount of information supplied by the new data are warranted.

Assessment of Predictive Uncertainty

Both parametric and conceptual model uncertainty is assessed in this study. Parameter estimates obtained from the least-square (or maximum likelihood) inverse modeling are subject to uncertainty, because of insufficient data and data error. The parameter estimation uncertainty is measured by the parameter estimation covariance matrix (Hill and Tiedeman 2007):

$$\boldsymbol{V}(\hat{\boldsymbol{\theta}}) = s^2(\boldsymbol{X}^T\boldsymbol{\omega}\boldsymbol{X})^{-1} \tag{8}$$

where $\boldsymbol{X}$ is sensitivity of matrix evaluated at the parameter estimate, $\hat{\boldsymbol{\theta}}$, and s^2 is calculated error variance

$$s^2 = \frac{\boldsymbol{e}^T\boldsymbol{\omega}\boldsymbol{e}}{N - N_k} \tag{9}$$

The covariance matrix, given as a byproduct of model calibration in MODFLOW-2000, is different for different models. According to the maximum likelihood theory (Berger 1985), the parameter estimates follow a multivariate normal distribution, whose mean and covariance are the calibrated parameters, $\hat{\boldsymbol{\theta}}$, and the covariance matrix $\boldsymbol{V}(\hat{\boldsymbol{\theta}})$, respectively. This allows generating random parameters for Monte Carlo simulations to assess the parametric estimation uncertainty of each model. After calculating the mean and standard deviation of head and flow rate of each model, the model averaging is implemented using Equations 3 and 4 to estimate the posterior mean and standard deviation.

Figures 9a to 9c plot the mean head simulations in northern Yucca Flat given by the averaged model and by

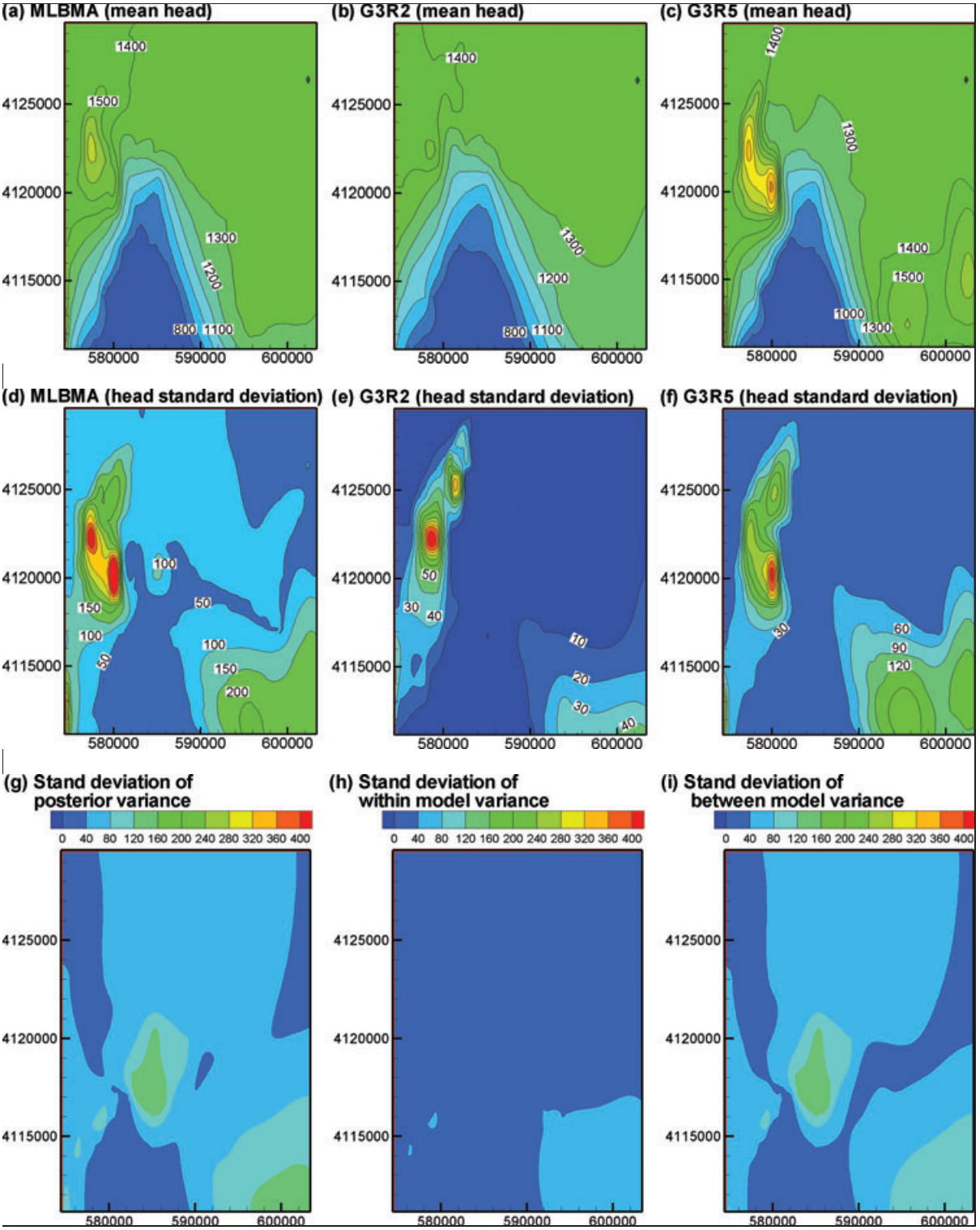

Figure 9. Mean and standard deviation of head simulations of (a and d) model averaging and models (b and e) G3R2 and (c and f) G3R5. (g to i) are for the standard deviation of (g) posterior, (h) within-model, and (i) between-model variance of head simulations. G3 is the CPT alternative model. R2 and R5 are NIM1 and CMB2 recharge models, respectively. The x- and z-coordinates are Easting and Northing (meters, UTM Zone 11, NAD 27).

models G3R2 and G3R5. A similar plot of the five most plausible models (having a sum of GLUE probabilities of 48.9%) is given as supporting information. The plotted results are for the final stress period and the top layer of the simulation domain. Although the general spatial patterns of the mean heads are similar for the individual models (high heads in the north and low heads in the south), significant differences in mean heads are observed between models (Figures 9b and 9c). The spatial patterns of individual models are preserved in the posterior mean

heads of the averaged model (Figure 9a) to different extents, depending on individual model probabilities. For example, the high mean heads simulated in the southeast corner by models G3R5 are not reflected in the posterior mean heads of the averaged model, because this model has a relatively lower probability. This aspect of model averaging is expected to avoid biased predictions that might result from using a single model.

Figures 9d to 9f are plotted in the same manner as Figures 9a to 9c, but for the standard deviations of heads simulated by model averaging and by the two individual models. The figure shows that, unlike the posterior mean of the averaged model, the spatial pattern of the posterior standard deviation is significantly different from that of individual models, and that the posterior standard deviation is larger throughout the entire area. This results from between-model variance (Equation 4) that is incorporated when averaging multiple models. As shown in Figures 9g to 9i, the spatial pattern and magnitude of the standard deviation of posterior head variance are dominated by those of the between-model variance, as compared with the within-model variance. This demonstrates that model averaging avoided underestimation of the magnitude and spatial distribution of the predictive uncertainty. The same conclusions can be drawn for the predictive uncertainty of flow rate, and the results and discussions are referred to Pohlmann et al. (2007).

Concluding Remarks

This paper evaluates two sources of model uncertainty for groundwater flow modeling in northern Yucca Flat of the DVRFS. Uncertainty in the recharge and geological model components results from different techniques for recharge estimation and different interpretations of geological and geophysical data, respectively. The uncertainty assessment is conducted in a flow-modeling framework by replacing the recharge and geological components of the DVRFS model with the alternative components. This results in 25 groundwater model combinations. These models are calibrated against the observations distributed over the entire DVRFS, and the calibration results in northern Yucca Flat are used for the uncertainty assessment. All four model selection criteria (AIC, AICc, BIC, and KIC) select two models as the most plausible and discard the other 23. However, using only two models is not considered adequate for this study because the 25 model predictions are significantly different and elimination of most of the alternative models may underestimate predictive uncertainty and bias the predictions. Instead, GLUE probabilities are used as they are more evenly distributed across all of the 25 models. The model averaging results are superior to those of individual models, as the former avoids the potential for biased predictions and underestimation of uncertainty. The magnitude and spatial distribution of the posterior variance is dominated by the between-model variance, which is significantly larger than the within-model variance corresponding to parametric uncertainty.

Uncertainty in the geological interpretations has a more significant effect on groundwater flow simulations than that in the recharge estimation methods. Different geological interpretations result in dramatically different flow patterns, whereas the effect of the recharge estimation methods is restricted to the top layers of the domain. For example, the different configurations of the CP thrust in geological interpretations G2 and G3 give entirely different flow patterns from the top to the bottom of the simulation domain. In addition, given a single recharge estimation method, the weighted residuals vary significantly from one geological model to another; on the contrary, the variation between multiple recharge estimation methods for a given geological interpretation is smaller. Most observations cannot be used to effectively discriminate between the alternative groundwater models because their residuals are similar in all models. Instead, the difference of model fit (measured by the SSWR) between different models is caused by a few observations (e.g., five head observations in this study). This finding is important for guiding data collection for further evaluation and reduction of the model uncertainty because it is more efficient to target areas where data collection will most effectively discriminate between alternative models.

Although GLUE probabilities are used here for model averaging, this does not suggest that GLUE is theoretically superior to the model selection criteria. Comparing predictive performance of GLUE and each model selection criteria requires running cross-validation (Ye et al. 2004; Foglia et al. 2007) for real world models, which is beyond the scope of this study. Real world models clearly present significant challenges to existing techniques for model averaging. One important issue is how to evaluate the model selection criteria. It is unknown whether the theoretical bases of the criteria are still valid if calibration data are different from the data used for calculating the criteria. In addition, it is still an open question as to whether one can select one or two models from several alternatives, on the basis of statistical criteria, for a site with complicated hydrogeologic conditions but sparse observations. Another issue is how model probabilities developed during flow modeling apply to subsequent transport modeling. This problem is resolved empirically in Pohlmann et al. (2007) and Reeves et al. (2009). Finally, the relative magnitudes of prior and posterior model probabilities are inconsistent. Although this is not surprising, the effect of using the informative prior probability on model predictions has not been quantified.

Acknowledgments

This research was supported in part by the U.S. Department of Energy, National Nuclear Security Administration Nevada Site Office under contract DE-AC52-00NV13609 with the Desert Research Institute (DRI). The

first author conducted part of the research when he was employed by DRI. The first author is also supported in part by NSF-EAR grant 0911074.

Supporting Information

Supporting Information may be found in the online version of this article:

The supporting information includes the definitions of the four model selection criteria, the procedure of developing the 25 alternative groundwater models and of assessing the model uncertainty, the reduction of SSWR in the new model calibration, and plots of mean and standard deviation of head simulations of MLBMA and the five most plausible models.

Table S1. The sum of squared weighted residuals (SSWR) of the four types of observations for the USGS geological model (G1) and the five recharge models. *N* is the number of observations of each type, and DVRFS is the SSWR of the original DVRFS model reported by Belcher et al. (2004). The observations are distributed throughout the entire DVRFS.

Table S2. The sum of squared weighted residuals (SSWR) of the two types of observations for the five geological and the five recharge models. *N* is the number of observations of each type, and DVRFS is the SSWR of the original DVRFS model reported by Belcher et al. (2004). The observations are distributed within the northern Yucca Flat.

Figure S1. Mean heads of (a) model averaging and models (b) G3R2, (c) G2R2, (d) G3R5, (e) G2R5, and (f) G2R3. G2 and G3 are the UGTA base model and its CPT alternative. R2, R3, and R5 are NIM1, NIM2, and CMB2 recharge models, respectively. The *x*- and *z*-coordinates are Easting and Northing (meters, UTM Zone 11, NAD 27).

Figure S2. Standard deviation of head predictions of (a) model averaging and models (b) G3R2, (c) G2R2, (d) G3R5, (e) G2R5, and (f) G2R3. G2 and G3 are the UGTA base model and its CPT alternative. R2, R3, and R5 are NIM1, NIM2, and CMB2 recharge models, respectively. The *x*- and *z*-coordinates are Easting and Northing (meters, UTM Zone 11, NAD 27).

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

References

Akaike, H. 1974. A new look at statistical model identification. *IEEE Transactions on Automatic Control* AC-19, 716–722.

Bechtel Nevada. 2006. *A Hydrostratigraphic Model and Alternatives for the Groundwater Flow and Contaminant Transport Model of Corrective Action Unit 97*. Nevada: Yucca Flat-Climax Mine, Lincoln and Nye Counties. DOE/NV/11718-1119.

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. Reston, Virginia: USGS.

Berger, J.O. 1985. *Statistical Decision Theory and Bayesian Analysis*, 2nd ed. New York: Springer-Verlag.

Beven, K. 2006. A manifesto for the equifinality thesis. *Journal of Hydrology* 320, 18–36.

Beven, K., and J. Freer. 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modeling of complex environmental systems using the GLUE methodology. *Journal of Hydrology* 249, no. 1-4: 11–29. DOI:10.1016/S0022-1694(01)00421-8.

Beven, K., and A. Binley. 1992. The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes* 6, no. 5: 279–283. DOI:10.1002/hyp.3360060305.

Bredehoeft, J.D. 2005. The conceptualization model problem—surprise. *Hydrogeology Journal* 13, 37–46.

Bredehoeft, J.D. 2003. From models to performance assessment: The conceptualization problem. *Ground Water* 41, no. 5: 571–577.

Carrera, J., and S.P. Neuman. 1986. Estimation of aquifer parameters under transient and steady state conditions: 3. Application to synthetic and field data. *Water Resources Research* 22, no. 2: 228–242.

Epstein, B. 2004. Development and uncertainty analysis of empirical recharge prediction models for Nevada's Desert basins. Masters thesis, University of Nevada. Reston, Virginia: USGS.

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, no. 5: 627–641.

Harbaugh, A.W., Banta, E.R., Hill, M.C., and McDonald, M.G. 2000. MODFLOW-2000, the U.S. Geological Survey modular ground-water model—User guide to modularization concepts and the ground-water flow process. U.S. Geological Survey Open-File Report 00-92. Reston, Virginia: USGS.

Harrar, W.G., T.O. Sonnenborg, and H.J. Henriksen. 2003. Capture zone, travel time, and solute-transport predictions using inverse modeling and different geological models. *Hydrogeology Journal* 11, 536–548.

Hevesi, J.A., A.L. Flint, and L.E. Flint. 2003. Simulation of net infiltration and potential recharge using a distributed parameter watershed model for the Death Valley Region, Nevada and California. Water Resources Investigations Report 03-4090. Sacramento, California: U.S. Geological Survey.

Hill, M.C., and C.R. Tiedeman. 2007. *Effective Methods and Guidelines for Groundwater Model Calibration, Including Analysis of Data, Sensitivities, Predictions, and Uncertainty*. Hoboken, New Jersey: John Wiley and Sons.

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. USGS Open-File Report 00-184. Reston, Virginia: USGS.

Hoeting, J.A., D. Madigan, A.E. Raftery, and C.T. Volinsky. 1999. Bayesian model averaging: A tutorial. *Statistical Science* 14, no. 4: 382–417.

Hojberg, A.L., and J.C. Refsgaard. 2005. Model uncertainty—parametric uncertainty versus conceptual models. *Water Science and Technology* 25, no. 6: 153–160.

Hurvich, C.M., and C.-L. Tsai. 1989. Regression and time series model selection in small sample. *Biometrika* 76, no. 2: 99–104.

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.

Maxey, G.B., and T.E. Eakin. 1949. Groundwater in White River Valley, White Pine, Nye, and Lincoln Counties, Nevada. No. 8, State of Nevada Office of the State Engineer prepared in cooperation with the U.S. Department of the Interior Geological Survey, Carson City, Nevada.

Neuman, S.P. 2003. Maximum likelihood Bayesian averaging of alternative conceptual-mathematical models. *Stochastic Environmental Research and Risk Assessment* 17, no. 5: 291–305. DOI: 10.1007/s00477-003-0151-7.

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.

Poeter, E.P., and M.C Hill. 2007. *MMA: A Computer Code for Multi-Model Analysis*. U.S. Geological Survey Techniques and Methods TM6-E3. Reston, Virginia: USGS.

Poeter, E., and D.R. Anderson. 2005. Multimodel ranking and inference in groundwater modeling. *Ground Water* 43, no. 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. Las Vegas, Nevada: U.S. Department of Energy.

Reeves, D.M., K.F. Pohlmann,, J.B. Chapman,, G.M. Pohll, and M. Ye. 2009. Influence of conceptual and parametric uncertainty on radionuclide flux estimates from a fracture granite rock mass. *Stochastic Environmental Research and Risk Analysis*, Under Review.

Refsgaard, J.C., J.P. van der Sluijs, A.L. Hojberg, and P.A. Vanrolleghem. 2007. Uncertainty in the environmental modeling process—a framework and guidance. *Environmental Modeling and Software* 22, no. 11: 1543–1556.

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. *Advances in Water Resources* 29, 1586–1597.

Rojas, R., L. Feyen, and A. Dassargues. 2009. Sensitivity analysis of prior model probabilities and the value of prior knowledge in the assessment of conceptual model uncertainty in groundwater modeling. *Hydrological Processes* 23, no. 8: 1131–1146.

Rojas, R., L. Feyen, and A. Dassargues. 2008. Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging. *Water Resources Research* 44, W12418. DOI:10.1029/2008WR006908.

Russell, C.E., and T. Minor. 2002. *Reconnaissance Estimates of Recharge Based on an Elevation-dependent Chloride Mass-balance Approach*. DOE/NV/11508-37, Publication No. 45164. Prepared for the U.S. Department of Energy, National Nuclear Security Administration Nevada Operations Office. Las Vegas, Nevada: Desert Research Institute.

Samper, F.J., and S.P. Neuman. 1989. Estimation of spatial covariance structures by adjoint state maximum likelihood cross-validation: 2. Synthetic experiments. *Water Resources Research* 25, no. 3: 363–371.

Scanlon, B.R., R.W. Healy, and P.G. Cook. 2002. Choosing appropriate techniques for quantifying groundwater recharge. *Hydrogeology Journal* 10, 18–39. DOI: 10.1007/s10040-001-0176-2.

Schwarz, G. 1978. Estimating the dimension of a model. *Annual Statistics* 6, no. 2: 461–464.

Sun, N.-Z., and W. W.-G. Yeh. 1985. Identification of parameter structure in groundwater inverse problem. *Water Resources Research* 21, no. 6: 869–883.

Troldborg, L., J.C. Refsgaard, K.H. Jensen, and P. Engesgaard. 2007. The importance of alternative conceptual models for simulation of concentrations in a multi-aquifer system. *Hydrogeology Journal* 15, 843–860.

Tsai F. T.-C., and X. Li. 2008. Multiple parameterization for hydraulic conductivity identification. *Ground Water* 46, no. 6: 851–864.

Tsai F. T.-C., N.Z. Sun, and W. W.-G. Yeh. 2003. A combinatorial optimization scheme for parameter structure identification in ground water modeling. *Ground Water* 41, no. 2: 156–169.

Ye, M., K.F. Pohlmann, and J.B. Chapman. 2008a. Expert elicitation of recharge model probabilities for the Death Valley regional flow system. *Journal of Hydrology* 354, 102–115. DOI:10.1016/j.jhydrol.2008.03.001.

Ye, M., P.D. Meyer, and S.P. Neuman. 2008b. On model selection criteria in multimodel analysis. *Water Resources Research* 44, no. 3: W03428. DOI:10.1029/2008WR006803.

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 Resources Research* 41, no. 12: W12429. DOI:10.1029/2005WR004260.

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. DOI:10.1029/2003WR002557.