Journal of Hydrology (2008) 354, 102–115

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jhydrol

Expert elicitation of recharge model probabilities for the Death Valley regional flow system

Ming Ye [a,*], Karl F. Pohlmann [b], Jenny B. Chapman [b]

[a] *School of Computational Science and Department of Geologic Sciences, Florida State University, Tallahassee, FL 32306, USA*
[b] *Desert Research Institute, Nevada System of Higher Education, 755 East Flamingo Road, Las Vegas, NV 89119, USA*

Received 12 January 2008; received in revised form 29 February 2008; accepted 3 March 2008

KEYWORDS
Model uncertainty;
Prior model probability;
Model averaging;
Expert elicitation;
Recharge estimates;
Death Valley regional flow system

Summary This study uses expert elicitation to evaluate and select five alternative recharge models developed for the Death Valley regional flow system (DVRFS), covering southeast Nevada and the Death Valley area of California, USA. The five models were developed based on three independent techniques: an empirical approach, an approach based on unsaturated-zone studies and an approach based on saturated-zone studies. It is uncertain which recharge model (or models) should be used as input for groundwater models simulating flow and contaminant transport within the DVRFS. An expert elicitation was used to evaluate and select the recharge models and to determine prior model probabilities used for assessing model uncertainty. The probabilities were aggregated using simple averaging and iterative methods, with the latter method also considering between-expert variability. The most favorable model, on average, is the most complicated model that comprehensively incorporates processes controlling net infiltration and potential recharge. The simplest model, and the most widely used, received the second highest prior probability. The aggregated prior probabilities are close to the neutral choice that treats the five models as equally likely. Thus, there is no support for selecting a single model and discarding others, based on prior information and expert judgment. This reflects the inherent uncertainty in the recharge models. If a set of prior probability from a single expert is of more interest, we suggest selecting the set of the minimum Shannon's entropy. The minimum entropy implies the smallest amount of uncertainty and the largest amount of information used to evaluate the models. However, when enough data are available, we prefer to use a cross-validation method to select the best set of prior model probabilities that gives the best predictive performance.
© 2008 Elsevier B.V. All rights reserved.

* Corresponding author. Tel.: +1 850 644 4587.
E-mail address: mingye@scs.fsu.edu (M. Ye).

0022-1694/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2008.03.001

Introduction

Uncertainty analysis of hydrologic models is an essential element for decision-making in water resource management. This paper is focused on conceptual model uncertainty, which arises when multiple conceptualizations of a hydrologic system (or its processes) are all acceptable given available knowledge and data. A model averaging concept has been developed to assess the conceptual model uncertainty by averaging predictions of multiple models using appropriate weights associated with each model. The weights can be calculated using likelihood functions (Beven, 2006 and its references therein) in the chi-square sense, the information criterion of AIC (Akaike, 1974) or AICc (Hurvich and Tsai, 1989) in the Kullback–Leibler sense (Burnham and Anderson, 2002, 2004; Poeter and Anderson, 2005), or the information criterion of BIC (Schwarz, 1978) or KIC (Kashyap, 1982) in the Bayesian sense (Draper, 1995; Hoeting et al., 1999; Neuman, 2003; Ye et al., 2004, 2005, 2008; Vrugt et al., 2006; Vrugt and Robinson, 2007). This paper addresses conceptual model uncertainty and model averaging in the Bayesian context.

In Bayesian model averaging (BMA) (Draper, 1995; Hoeting et al., 1999) or its maximum likelihood version (MLBMA) (Neuman, 2003), if Δ is a quantity that one wants to predict, then its posterior distribution given conditioning data $\mathbf{D}$ (including measurements of model parameters and observations of state variables) is the average of the distributions $p(\Delta|M_k,\mathbf{D})$ under each model M_k weighted by the posterior model probability $p(M_k|\mathbf{D})$, i.e.,

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

The posterior model probability, $p(M_k|\mathbf{D})$, is estimated via the Bayes' theorem

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

where $p(\mathbf{D}|M_k)$ is the model likelihood function and can be approximated by $p(\mathbf{D}|M_k) = \exp(-\mathrm{KIC}_k/2)$ or $p(\mathbf{D}|M_k) = \exp(-\mathrm{BIC}_k/2)$ (Ye et al., 2004), and $p(M_k)$ is prior probability of model M_k. Summation of the prior probabilities of all the alternative models is one,

$$\sum_{k=1}^{K} p(M_k) = 1 \quad (3)$$

implying that all possible models of potential relevance to the problem at hand are under study, and that all models differ from each other sufficiently to be considered mutually exclusive (the joint probability of two or more models being zero). The question of how to assign prior probabilities $p(M_k)$ to models M_k remains largely open. A common practice is to adopt a ''reasonable 'neutral' choice'' (Hoeting et al., 1999), according to which all models are initially considered to be equally likely, there being insufficient prior reason to prefer one over another. However, the neutral choice of prior model probabilities ignores expert knowledge of the system to be modeled, thereby implying maximum ignorance on the part of the analyst.

Generally speaking, the prior model probability is an analyst's (or a group of analysts') subjective degree of reasonable belief (Jeffreys, 1957) or confidence (Zio and Apostolakis, 1996) in a model. The belief or confidence is ideally based on expert judgment. Using expert judgments is prevalent in uncertainty and risk analysis (Cooke, 1991; Ayyub, 2001; Bedford et al., 2006), especially when experimental and statistical evidence is insufficient (Refsgaard et al., 2006). For a complicated hydrologic system, expert judgment or experience is the basis of conceptual model development, and may be more informative than limited observations. This is particularly true for subsurface hydrology, where hydraulic parameters are measured from sparse samples (boreholes) and mathematical models may disagree with geologic rules (Wingle and Poeter, 1993; Lele and Das, 2000). Garthwaite et al. (2005) argue that a better use of expert judgment could add more information than slight improvement of data analysis techniques.

Hence, we view integrating expert judgment in BMA (by specifying subjective prior probabilities) to be a strength rather than a weakness. Madigan et al. (1995) and Zio and Apostolakis (1996) demonstrated that using informative prior model probabilities (in contrast to equal ones) on the basis of expert judgment can improve model simulation and uncertainty assessment. Ye et al. (2005) developed a constrained maximum entropy method, which estimates informative prior model probabilities through the maximization of the Shannon's entropy (Shannon, 1948) subject to constraints reflecting a single analyst's (or group of analysts') prior perception about how plausible each alternative model (or a group of models) is relative to others, and selection of the most likely among such maxima corresponding to alternative perceptions of various analysts (or groups of analysts). By running cross-validation, Ye et al. (2005) demonstrated that, in comparison to using equal prior model probabilities, using informative probabilities improves model predictive performance.

The subjective prior model probabilities can be directly obtained through expert elicitation. The expert elicitation has been applied to many studies, for example, future climate change (Arnell et al., 2005; Miklas et al., 1995), performance assessment of proposed nuclear waste repositories (Hora and Jensen, 2005; McKenna et al., 2003; Draper et al., 1999; Hora and von Winterfeldt, 1997; Zio and Apostolakis, 1996; Morgan and Keith, 1995; DeWispelare et al., 1995; Bonano and Apostolakis, 1991; Bonano et al., 1990), estimation of parameter distributions (Parent and Bernier, 2003; Geomatrix Consultants, 1998; O'Hagan, 1998), development of Bayesian network (Pike, 2004; Stiber et al., 1999, 2004; Ghabayen et al., 2006), and interpretation of seismic images (Bond et al., 2007). Formal expert elicitation processes have been proposed by Hora and Iman (1989) and Keeney and von Winterfeldt (1991), among others. Although expert elicitation is criticized in various aspects, such as selection of experts and accurate expression of experts' knowledge and belief in probability forms (O'Hagan and Oakley, 2004), the quality of educing expert judgments can be controlled by a formal procedure of expert elicitation and documentation (Garthwaite et al., 2005). Nevertheless, expert judgments should be used with caution, not to replace ''hard'' science (Apostolakis, 1990). When assessing conceptual model uncertainty, it is essential to adjust the prior probability to obtain the posterior model probability by conditioning of on-site measurements and observations.

Different from general uses of expert elicitation for model parameterization and development, this paper uses the expert elicitation to estimate prior model probabilities of alternative models. With few examples of such an application of expert elicitation in model uncertainty assessment (Zio and Apostolakis, 1996; Draper et al., 1999; Curtis and Wood, 2004), this study is expected to provide theoretical and practical guidelines for future applications of expert elicitation. This paper is focused on development of prior model probabilities using expert elicitation; discussion of using on-site data to further evaluate the alternative models is beyond our scope here.

The expert elicitation is used in this paper to estimate prior probabilities of five recharge models developed for the Death Valley regional flow system (DVRFS), covering southwestern Nevada and the Death Valley area of eastern California, USA (Fig. 1a). Due to existing and potential radionuclide contamination at the US Department of Energy's Nevada Test Site (NTS) and the proposed Yucca Mountain high-level nuclear waste repository in the DVRFS, it is critical to predict contaminant transport in the region. Hydrologic and geologic conditions in the DVRFS are complicated, rendering multiple conceptualizations of the system based on limited data and information. Because conceptual model uncertainty can be significant, ignoring it (focusing only on parametric uncertainty) may result in biased predictions and underestimation of uncertainty. While expert elicitation was used for evaluating uncertainty of recharge and geological models (Pohlmann et al., 2007), this paper focuses on the recharge models applied throughout the DVRFS. In the past few decades, several recharge models have been independently developed for Nevada by different researchers based on different scientific theories. These include the Maxey–Eakin model (Maxey and Eakin, 1949), the discrete-state compartment model (Kirk and Campana, 1990; Carroll et al., 2007), the elevation-dependent chloride mass balance model (Russell and Minor, 2002; Russell, 2004; Minor et al., 2007) and the distributed parameter watershed model (Hevesi et al., 2003). It is unclear to scientists working in the DVRFS which recharge model should be used for groundwater flow and contaminant transport modeling. As recharge is the major driving force of groundwater flow, and thus contaminant transport, in the arid environment of the DVRFS, it is important to understand recharge model uncertainty. Our ultimate goal is to incorporate the recharge model uncertainty in our uncertainty analysis of DVRFS groundwater models.

It is worth pointing out that recharge model uncertainty is prevalent and not limited to the DVRFS. Recharge is a fundamental component of groundwater systems, and with

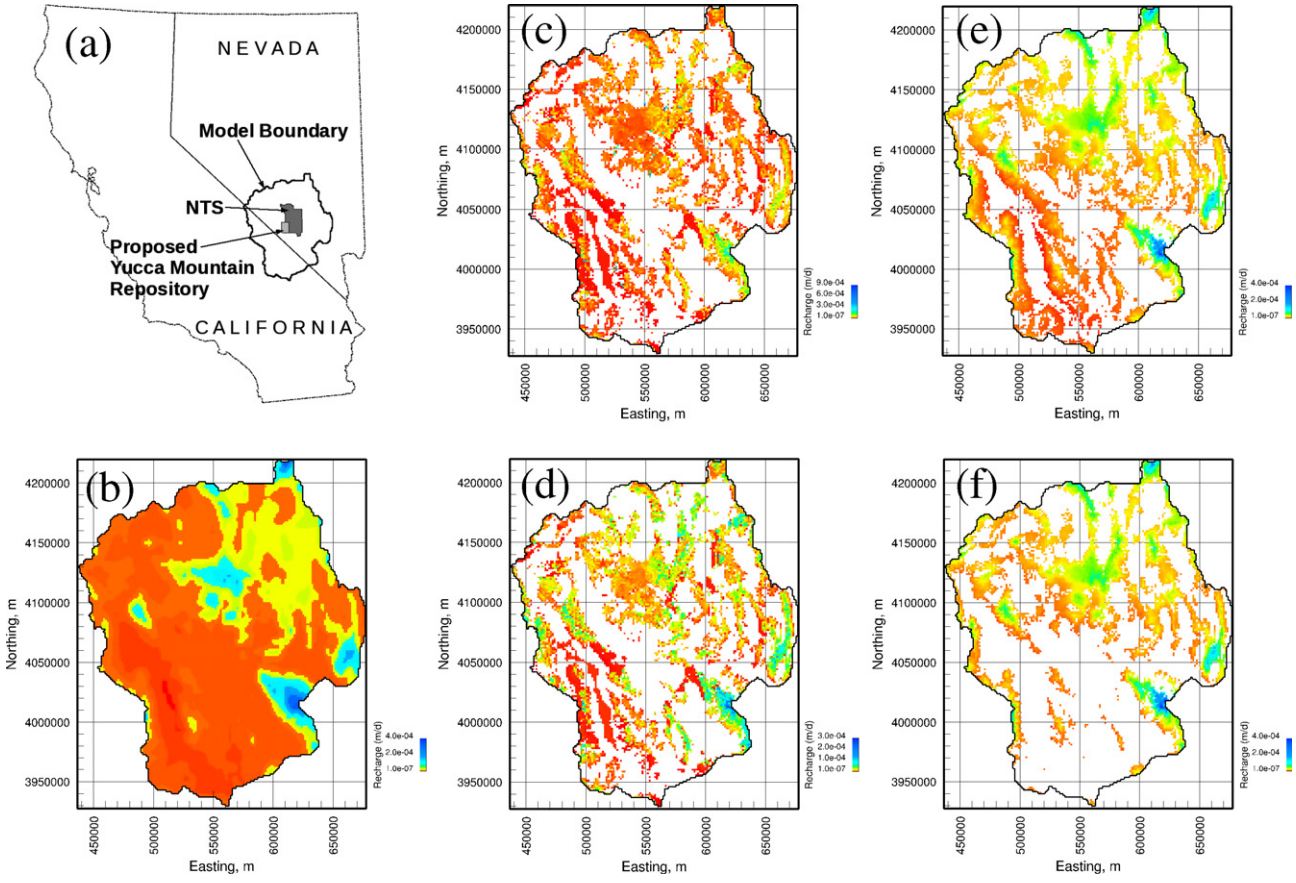

Figure 1 (a) Boundaries of the Death Valley regional flow system, the Nevada Test Site, the proposed Yucca Mountain nuclear waste repository, and recharge rate estimates (m/d) of models (b) MME (modified Maxey–Eakin model), (c) NIM1 (net infiltration model with runon–runoff component), (d) NIM2 (net infiltration model without runon–runoff component), (e) CMB1 (chloride mass balance model with alluvial mask), and (f) CMB2 (chloride mass balance model with alluvial and elevation masks).

multiple recharge estimation methods (or models) available, it is nontrivial to select the recharge estimation method appropriate for a given environment (see review articles of Scanlon et al., 2002; Scanlon, 2004). Scanlon et al. (2002) suggested using multiple methods to enhance reliability of recharge estimates. This is in line with the new concept of model averaging discussed above.

The second section of this paper introduces the recharge models considered in the expert elicitation. Recharge estimates of the models are briefly compared in terms of their values, spatial distributions and statistical characteristics. In particular, we explain the reasons for treating recharge uncertainty as conceptual model uncertainty, rather than as parametric uncertainty. The process of expert elicitation is listed in the third section, followed by discussion of elicitation results in the fourth section. Our conclusions are summarized in the fifth section.

Description of the five alternative recharge models

The five recharge models considered for the DVRFS are described briefly below; details of the models can be found in their original publications. Additional comparison of the models can be found in Rehfeldt (2004) and Pohlmann et al. (2007). Description of the geologic, hydrologic and hydrogeologic conditions of the DVRFS is beyond the scope of this paper, and the reader is referred to D'Agnese et al. (1997) and Belcher (2004) for further information on these topics.

Modification of the Maxey—Eakin method (MME)

Maxey and Eakin (1949) presented an empirical method (known as the Maxey—Eakin method) for estimating groundwater recharge as a function of precipitation. Since its inception, the Maxey—Eakin method has become the predominant technique used for estimating annual groundwater recharge in Nevada. The method estimates recharge via

$$R = \sum_{i=1}^{N} C_i P_i \quad (4)$$

where R is the estimated recharge, C_i are the percentage adjustment coefficients, P_i are the annual precipitation values within zones of precipitation and N is the number of precipitation zones. Maxey and Eakin (1949) utilized the precipitation map for Nevada developed by Hardman (1936) that includes hand-drawn contours based on weather station records and topography. The precipitation is distributed among five isohyets ($N = 5$) of 5, 8, 12, 15 and 20 in. Assuming a steady-state basin flow condition in which discharge from a basin is approximately the same as recharge into the basin, the coefficients, C_i, were developed through a trial-and-error method to attain a general agreement between the volumes of estimated recharge and measured discharge for 13 basins in eastern and central Nevada. The coefficients, listed in Table 1, increase in magnitude as the amount of precipitation increases while evapotranspiration and surface water runoff presumably decline. Note that the precipitation zone receiving less than 8 in./yr rainfall does not contribute to groundwater recharge.

Given the incomplete coverage of the DVRFS domain by the Hardman precipitation map, Epstein (2004) modified the Maxey—Eakin model, hereinafter referred to as the modified Maxey—Eakin model (MME). The method uses the PRISM map (Precipitation Estimation on Independent Slopes Model) (Daly et al., 1994) so that the recharge is estimated in a consistent way over both the Nevada and California portions of the DVRFS. Considering uncertainty in the PRISM estimates of precipitation, the MME evaluates uncertainty of the recharge coefficients, C_i, using an automated calibration method based on 91 basins. Table 1 lists the mean coefficients of four precipitation zones (thus $N = 4$ in MME) used to estimate recharge of the DVRFS. Different from the Maxey—Eakin method, the coefficient for the lowermost precipitation zone is allowed to be nonzero. Although the MME model is more complicated than the original ME model, it is still the simplest model in the model set. The recharge map of the DVRFS estimated using the MME (with the mean coefficients) is shown in Fig. 1b.

Two net infiltration models (NIM)

Hevesi et al. (2003) developed a distributed-parameter watershed model, INFILv3, for estimating temporal and spatial distribution of net infiltration and potential recharge in the Death Valley region, including the DVRFS. The estimates of net infiltration quantify downward drainage of water across the lower boundary of the root zone, and are used as an indication of potential recharge under current climate conditions. Based on the daily average water balance at the root zone, the model comprehensively represents processes controlling net infiltration and potential recharge. The daily water balance includes the major components of the water balance for arid to semiarid environments, including precipitation; infiltration of rain; snowmelt and surface water into soil or bedrock; *runoff* (excess rainfall and snowmelt);

Table 1 Recharge coefficients for the Maxey—Eakin method and the modified Maxey—Eakin method (Epstein, 2004)

Maxey—Eakin method		Modified Maxey—Eakin method	
Precipitation zone (in./yr)	Coefficient	Precipitation zone (in./yr)	Coefficient
0 to less than 8	0.00	0 to less than 10	0.019
8 to less than 12	0.03	10 to less than 20	0.049
12 to less than 15	0.07	20 to less than 30	0.195
15 to less than 20	0.15	Greater than 30	0.629
Greater than 20	0.25		

surface water *runon* (overland flow and streamflow); bare-soil evaporation; transpiration from the root zone; redistribution or changes in water content in the root zone; and net infiltration across the lower boundary of the root zone. Various techniques were developed to estimate these quantities and their spatial and temporal variability, which renders this method comprehensive but complicated. The model parameters (e.g., bedrock and soil saturated hydraulic conductivity and root density) were adjusted through model calibration by comparing simulated and observed streamflow as well as basin-wide average net infiltration and previous estimates of basin-wide recharge.

Two alternative net infiltration models *with* and *without* runon–runoff component (Hevesi et al., 2003) are considered in this paper to represent the two opposite conceptualizations. Fig. 1c and d depicts the averaged annual net infiltration estimates of the two models. Groundwater recharge can be estimated from the net infiltration estimates by multiplying the net infiltration with coefficients related to rock hydraulic conductivity at the water table, since the net infiltration distribution only accounted for surficial characteristics of the system. For more details about the determination of the coefficients, the reader is referred to Belcher (2004). For convenience in this discussion, the two net infiltration models are also referred to as recharge models.

Two elevation-dependent chloride mass balance models (CMB)

The chloride mass balance (CMB) method estimates recharge in basins (or any hydrologic systems) based on a balance between chloride mass within hydrologic input and output components. The method assumes that chloride in groundwater within the basins originates from chloride in precipitation in mountain uplands and dry-fallout and is transported to adjacent valleys by steady-state groundwater flow (Dettinger, 1989). At its most fundamental level, the method requires only estimates of annual precipitation in the recharge areas, total chloride input (chloride concentrations in precipitation and recharge water) and total chloride output (chloride concentrations in adjacent basin groundwater). The rate of recharge, R, can be calculated as (Maurer et al., 1996)

$$R = \frac{C_p P}{C_r} - \frac{C_{sw} S_w}{C_r} \tag{5}$$

where C_p is the combined wet-fall and dry-fall atmospheric chloride concentration normalized to precipitation, P is the mean annual precipitation rate, C_r is the chloride concentration in recharge water and C_{sw} is the chloride concentration in surface water runoff S_w. For individual basins, recharge rate can be estimated from this information if the following assumptions are met (Dettinger, 1989): (1) there are no other major sources or sinks for chloride in the system; (2) surface runoff is small in comparison to groundwater flow; and (3) the recharge areas are correctly delineated. Russell and Minor (2002) extended the chloride mass balance approach to account for the elevation of precipitation, the limited quantities of recharge that are thought to occur on low-elevation alluvial surfaces, and uncertainty inherent in the data. This elevation-dependent chloride mass balance approach was applied by Russell and Minor (2002) to a 7900-km^2 region of the Nevada Test Site (NTS) and vicinity within the DVRFS.

Although this recharge/elevation relationship simulates recharge at all elevations, several studies suggest that significant groundwater recharge does not occur through low-elevation alluvial sediments in southern Nevada. Russell and Minor (2002) thus developed two models to address this uncertain conceptualization of low-elevation recharge. The first model assumes that all land surface areas covered by alluvial sediments receive negligible recharge based on the results of previous studies and soil-water chloride profiles of 40 boreholes completed in unsaturated alluvium within the NTS (Russell and Minor, 2002). This model is called the *CMB model with alluvial mask*. The second model assumes that the elevation of the lowest perennial spring that discharges from a perched groundwater system in the study area represents the lowest elevation at which significant recharge occurs. This spring is Cane Spring, which is located at an elevation of 1237 m above mean seas level. Coincidentally, this is approximately the same elevation (1200 m) that Harrill (1976) and Dettinger (1989) consider to be the minimum at which precipitation makes a significant contribution to recharge in desert basins of central and southern Nevada. Using the concept of a recharge cut-off elevation, Russell and Minor (2002) define a zone of zero recharge that encompasses all elevations below 1237 m plus elevations above 1237 m that are covered by alluvium. This model is called *CMB with both elevation and alluvial masks*. To assess uncertainty in the model parameters and measurements (e.g., precipitation and chloride concentration in spring water), Russell and Minor (2002) developed a Monte Carlo method to estimate multiple realizations of the recharge estimates. The two models were further extended in Russell (2004) and this study to include more basins in Nevada and cover the DVRFS. Fig. 1e and f depicts mean recharge estimates of the two CMB models.

Summary and discussion

The five recharge models are summarized as follows:

MME (Fig. 1b): modified Maxey–Eakin model using the mean coefficients.
NIM1 (Fig. 1c): net infiltration model with runon–runoff component.
NIM2 (Fig. 1d): net infiltration model without runon–runoff component.
CMB1 (Fig. 1e): chloride mass balance model with alluvial mask (mean estimates only).
CMB2 (Fig. 1f): chloride mass balance model with alluvial and elevation masks (mean estimates only).

Fig. 1 illustrates similarities and differences of the recharge rate estimates (m/d) of the five models, and Table 2 lists the total recharge estimates (m^3/d) for the entire DVRFS by each method. The MME gives the highest recharge estimate, and the CMB models give higher estimates than the NIM models. Due to the runon–runoff component considered in NIM1, the recharge estimate of NIM1 is higher than that of NIM2, while spatial patterns of the recharge estimate are similar in the two models. Because of the extra

Table 2 Recharge estimates (m^3/d) of the five recharge models in the DVRFS

Recharge model	DVRFS (m^3/d)
MME	596,190.8
NIM1	341,930.6
NIM2	282,223.1
CMB1	385,213.7
CMB2	365,647.2

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). The recharge estimate of the MME has the smoothest spatial distribution, due to the four precipitation zones. The different recharge estimates are viewed as a result of conceptual model uncertainty, rather than parametric uncertainty, since they are caused by simplification and inadequacy/ambiguity in describing the recharge process and not by uncertainty in recharge measurements themselves (Wagener and Gupta, 2005).

Given the five recharge models, which model (or models) should be used for groundwater modeling? Is it reasonable and justifiable to select a single model and to discard others based on expert judgment? How should uncertainty of the recharge models be assessed? The expert elicitation is used to answer these questions, and ultimate results of this expert elicitation are the prior model probabilities essential to the BMA for assessing the conceptual model uncertainty.

Process of the expert elicitation

While several processes of expert elicitation have been suggested in the literature (e.g., Hora and Iman, 1989; Bonano et al., 1990), the process proposed by Keeney and von Winterfeldt (1991) was followed, since it is closely pertinent to eliciting probability from experts and has been applied to model probability elicitation (Zio and Apostolakis, 1996). The formal process consists of the seven steps listed below. Implementation of the process for the recharge models is also described.

Step 1: Identification and selection of elicitation issues

The elicitation issues are the questions posed to the experts that require their answers. The following three issues are considered for assessing the recharge model uncertainty:

(1) *Is the model set complete, given the objective of the analysis?* BMA requires that alternative models are comprehensively exhaustive (all alternative models are included in the model set). Since this requirement cannot be satisfied in an absolute sense, we elicit from the experts whether there are other alternative models that are comparable in importance to the five models and should be considered.

(2) *What are the plausibility ranks of these models, given the objective of the analysis?* Whereas ranking of model plausibility is qualitative and the ranks cannot directly give the prior model probability, the model ranking helps experts evaluate relative plausibility of the models before they estimate prior model probability.

(3) *What is the probability value that best represents the confidence you would place on each recharge model, given the objective of the analysis?* Model probabilities are the ultimate goal of the expert elicitation, and will be used directly in the BMA to calculate the posterior model probability through Eq. (2).

Step 2: Identification and selection of experts

Expert elicitation requires three types of experts: generalists, specialists and normative experts. In this study, the generalists should be knowledgeable about various aspects of the recharge models and the broader study goals (in this case, assessing groundwater flow and contaminant transport in the DVRFS). They typically have substantive knowledge in one discipline (e.g., geology or hydrology) and a general understanding of the technical aspects of the problem. While the generalists are not necessarily at the forefront of any specialty within their main discipline, the specialists should be at the forefront of one specialty relevant to the recharge models. The specialists often do not have the generalists' knowledge about how their expertise contributes to the broader study with respect to recharge model uncertainty analysis. Normative experts typically have training in probability theory, psychology and decision analysis. They assist generalists and specialists in articulating their professional judgments and thoughts so that they can be used in a meaningful way in the conceptual model uncertainty assessment. A high-quality elicitation requires the teamwork of all three types of experts.

Selecting experts is a time-consuming process, and may take more than a year for a full-scale elicitation (e.g., having international nomination of experts and forming an expert panel of international scientists, as in Hora and Jensen, 2005). With practical limitations, we selected national and state experts, who were believed well-qualified owing to their familiarity with the hydrogeologic conditions of the DVRFS and their research at the forefront of recharge estimation in semi-arid environments of the southwestern US. Five specialists, two generalists and one normative expert were identified. The normative expert had an advisory role and was not involved in evaluating the recharge model uncertainty.

Step 3: Discussion and refinement of elicited issues

This step allows discussion and refinement, if necessary, of the issues and quantities that will be elicited. While Keeney and von Winterfeldt (1991) suggest completing this step by a 1-day meeting of all experts, such a meeting was considered unnecessary for this project. Instead, one month before the elicitation, the experts received the three clearly stated elicitation issues, as well as original publications of the five recharge models and references about conceptual model uncertainty, BMA, prior model probability and expert judgment. The experts studied these materials, and some discussed details of the models with us and requested more reading materials.

Step 4: Training for the elicitation

Led by the normative expert, the training was conduct in two meetings in the first half day of elicitation. In the first training meeting, the normative expert introduced the

three elicitation issues, the purpose of the broader study, the quantities to be elicited, the concept of BMA and its application in Ye et al. (2004), and the estimation of prior model probability in Ye et al. (2005). The example of Zio and Apostolakis (1996) using experts for estimating prior model probability was also introduced. It is critical to make clear to the experts that the probability is expressed, in a Bayesian point of view, as a subjective degree of belief. The second training meeting further familiarized the experts with the recharge models. Presentations of the recharge models were made to the experts, and a lively debate ensued among the experts about advantages, disadvantages, assumptions and the most appropriate application areas of these models.

During the first training meeting, the three types of biases that may occur during elicitation were introduced: overconfidence, anchoring and availability (Keeney and von Winterfeldt, 1991). ''Overconfidence'' is to express more certainty than is appropriate and assign a large prior probability to certain models. ''Anchoring'' is to hesitate to adjust the prior model probability but to focus on its initial value. ''Availability'' is to overemphasize the events that are easily imagined or recalled. Bias can also occur if experts focus on concrete evidence and data as a main source of probability judgments and ignore more abstract information. Bedford et al. (2006) listed two more biases: motivational, the situation where the expert is interested in a particular value, and cognitive, which concerns the situation where the expert incoherently gives an assessment based on a number of calculations.

Step 5: Elicitation

Since most of the elicitation can be completed within 1–3 h (Keeney and von Winterfeldt, 1991), the elicitation was conducted in the second half day of elicitation. All experts were asked to answer a questionnaire with seventeen questions. Since the questions may be useful as examples for other elicitations, they are listed in Appendix A. The questions were designed in the order that progressively quantitative questions follow qualitative ones. Assignment of prior model probabilities was the last question. Experts were also required to provide justifications for their answers.

Step 6: Analysis, aggregation and resolution of disagreement

Immediately after the elicitation, when the meeting was still vivid in memory, the experts' answers were analyzed and aggregated to yield the final estimation of the elicited quantity. Keeney and von Winterfeldt (1991) suggested resolving the disagreements between the answers by having a meeting after the elicitation. This was not considered necessary in our case, since different distributions of model plausibility reflect experts' different degree of belief regarding model uncertainty. Phrasing the uncertainty personally encourages the expert to provide his opinion without the burden of representing some broader consensus view. The expert elicitation of Bond et al. (2007) indicated that it is difficult to resolve the conceptual model uncertainty by consensus. In addition, since we aggregated the elicited model probabilities using a mathematical method (described below), not behavioral approaches, there is no need to arrive at a consensus distribution.

The simplest aggregation is the arithmetic mean of the elicited model probability via

$$p_i = \frac{1}{M}\sum_{k=1}^{M} p_{ki} \tag{6}$$

where M is the number of experts, and p_{ki} is the probability that expert k assigns to model i. Since simple averaging does not consider between-expert variability, we used an iterative aggregation method of De Groot (1974). This method requires each expert to assign averaging weights to his and other experts' judgments, $\mathbf{w} = [w_{ij}]$, where w_{ij} is the weight that expert i assigns to expert j and $\sum_j w_{ij} = 1$. The weight, also subjective, incorporates between-expert variability. The elicited prior probabilities are expressed as a matrix, $\mathbf{p} = [p_{ij}]$, where p_{ij} is the probability that expert i assigns to model j and $\sum_j p_{ij} = 1$. After ''learning'' the assessments from all the other experts, expert i could change his probability, on this view, to

$$\mathbf{p}' = \mathbf{wp}; \quad p'_{ij} = \sum_{k=1}^{M} w_{ik}p_{kj} \tag{7}$$

where M is the number of experts. There being no reason to stop at $\mathbf{p}'$, the expert could change again to

$$p'' = \mathbf{wp}' = \mathbf{w}(\mathbf{wp}) \tag{8}$$

This process converges to a matrix $\mathbf{w}^{\infty}\mathbf{p}$, where the rows of $\mathbf{w}^{\infty}$ are all the same. This indicates that, by iteratively ''revising'' their own opinions in the above manner, the experts all converge toward the same probability vector.

Step 7: Documentation and communication

The following material related to the elicitation process should be well documented: (1) elicitation issues and quantities, (2) expert identification and selection, (3) training material, (4) training and elicitation process and the results from each expert, (5) aggregation of the elicited quantities and (6) final model probabilities.

Results and discussion

Elicitation results for the recharge models are presented and discussed in this section. At the end of this section, a companion elicitation regarding alternative geologic models is compared and contrasted with the recharge elicitation. Details of the geological models and the corresponding elicitation process are presented by Pohlmann et al. (2007).

Model set completeness

While four experts considered that the recharge model set is complete, three experts suggested adding a tracer (deuterium) technique used together with the discrete-state compartment (DSC) method (Feeney et al., 1987; Kirk and Campana, 1990; Sadler et al., 1991). This model is based on saturated-zone studies and, in this sense, is similar to the chloride mass balance models included in the model set. It divides a flow domain into various cells (basins or sub-basins) and estimates uniform recharge within each cell by calibrating tracer mass estimates against site measurements. Because it does not incorporate recharge spatial variability within each large cell, this model does not provide recharge information at a scale consistent with the other recharge models and necessary for the DVRFS model. It is thus regarded as incomparable with the current five recharge

models and not considered further. If the DSC model were modified to provide information at the appropriate scale and included in the model set, it is likely it would have received the smallest model probability because of the extremely sparse isotopic data supporting it. Due to its similarity to the CMB models, the prior probabilities of the CMB models might have decreased accordingly.

Model plausibility ranking and prior model probability

Model rankings elicited from the experts are plotted in Fig. 2. The experts gave significantly different model rankings, reflecting the different individual perceptions of model plausibility. For example, Expert 1 ranked the NIM1 model as the most plausible, whereas Expert 2 ranked this model as the least plausible. Expert 2 suspected the reliability of inputs of the NIM models; however, four out of seven experts believed that the most complicated NIM1 model can give better estimates than other models. On average, the NIM1 and NIM2 models received the highest and lowest overall ranking, respectively. This is not surprising, since the models are based on two opposite (with and without) conceptualizations of the runon–runoff component. The CMB2 model received higher ranking than the CMB1 model, but lower than the MME model. It is interesting that no expert considered the MME model the least plausible, although the model gives relatively higher and coarser recharge estimates.

The model ranking is consistent with the elicited prior model probability plotted in Fig. 3. Although the experts evaluated the models from various aspects (e.g., model assumptions and calibration results), the figure shows that no model received more than 50% prior probability from any expert. This indicates that there is no support from the experts to select a single recharge model for groundwater modeling, though this is commonly done. The MME, NIM1 and CMB2 models are the three most plausible models; none of these received the smallest prior probability (5%) from any expert. These three models belong to three different recharge technique categories (Scanlon, 2004): empirical approach (MME), recharge approach based on unsaturated-water studies (NIM1) and recharge approach based on saturated-water studies (CMB2). The elicited prior probabilities (less than 50%) suggest that the bias of overconfidence did not occur, while it is unclear whether the bias of anchoring and availability occurred during the elicitation.

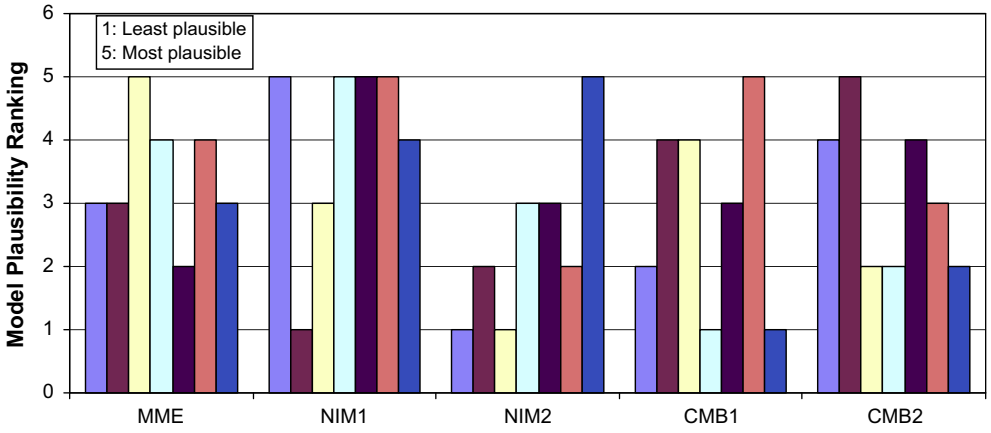

Figure 2 Column chart of the plausibility ranking of the five recharge models. The columns of each model represent elicited model ranking from the seven experts. The most plausible model is ranked 5 and the least plausible model is ranked 1.

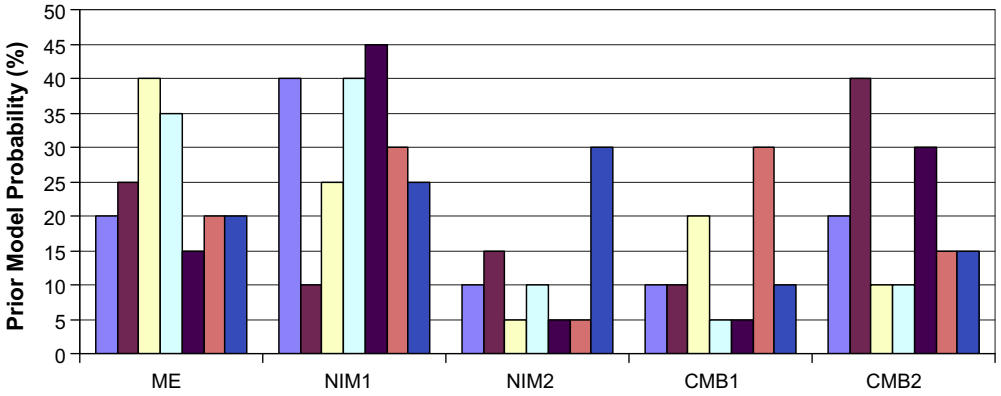

Figure 3 Column chart of the prior probability of the five recharge models. The columns of each model represent elicited prior model probability from the experts.

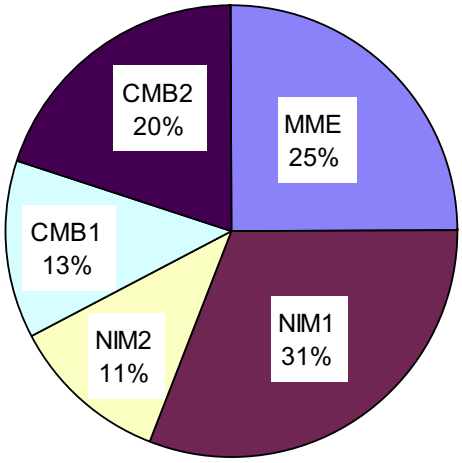

Figure 4 Aggregated prior probabilities from the simple averaging for the five recharge models.

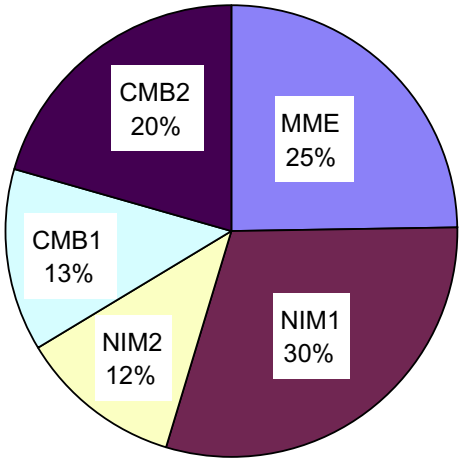

Figure 5 Iteratively aggregated prior probability of the five recharge models with consideration of expert-to-expert variability.

Aggregation of prior model probability and expert weight

Figs. 4 and 5 plot the aggregated prior model probabilities using the simple averaging and iterative methods, respectively. Three iterations were conducted to reach the final aggregation. The aggregated prior model probabilities of the two methods are almost identical, except for the 1% difference for the NIM1 and NIM2 models. The negligible difference results from the more-or-less uniform expert weights assigned by each expert. Fig. 6 shows that Experts 1, 2, 6 and 7 assigned the same or almost the same weights to all the experts. Despite the small difference between the two aggregation methods, the iterative aggregation method is still preferred, since it provides a formal way to reach consensus with consideration of expert-to-expert variability. Fig. 5 shows that the NIM1 and NIM2 models have the largest and smallest probability, respectively. Probability of CMB2 is larger than that of CMB1, but less than that of MME. This order of aggregated model probability is consistent with the model ranking and probabilities plotted in Figs. 2 and 3.

Although different models received significantly different prior probabilities from each expert (Fig. 3), the aggregated probabilities are more-or-less uniform, considering that equally likely prior probability is 20%. The largest deviation from the equally likely prior probability is only 10% for the NIM2 model. This manifests the inherent uncertainty in the recharge models. Given the final prior model probabilities, there is no justification to select one model and discard others based on prior information and expert judgment.

Discussion

Relatively uniform aggregated prior probabilities were also observed in another elicitation regarding five alternative geological models at the Climax Stock area of the DVRFS (Pohlmann et al., 2007). Uncertainty in some aspects of the geologic framework led to alternative interpretations of stratigraphic sequence and structure that could affect groundwater flow (e.g., a thrust fault could juxtapose an impermeable unit and an aquifer at depth). This elicitation

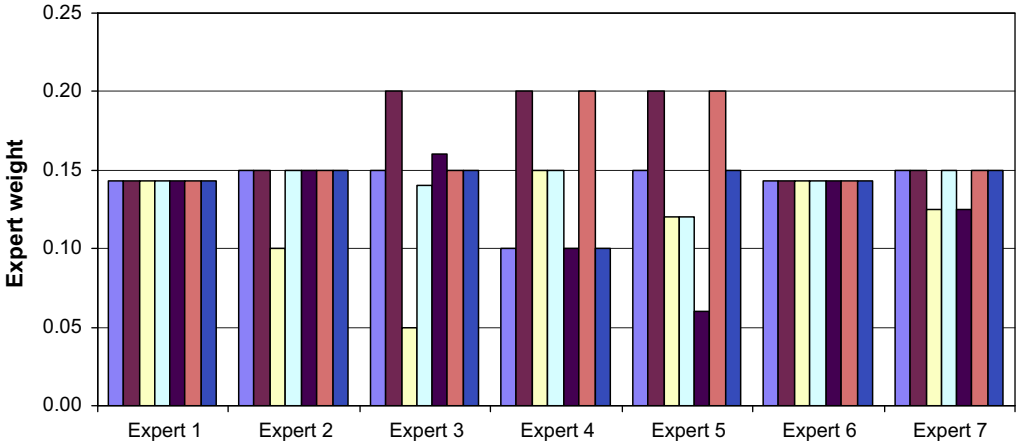

Figure 6 Expert weights assigned by each expert to all experts. The columns of each expert represent the weight assigned to all the experts.

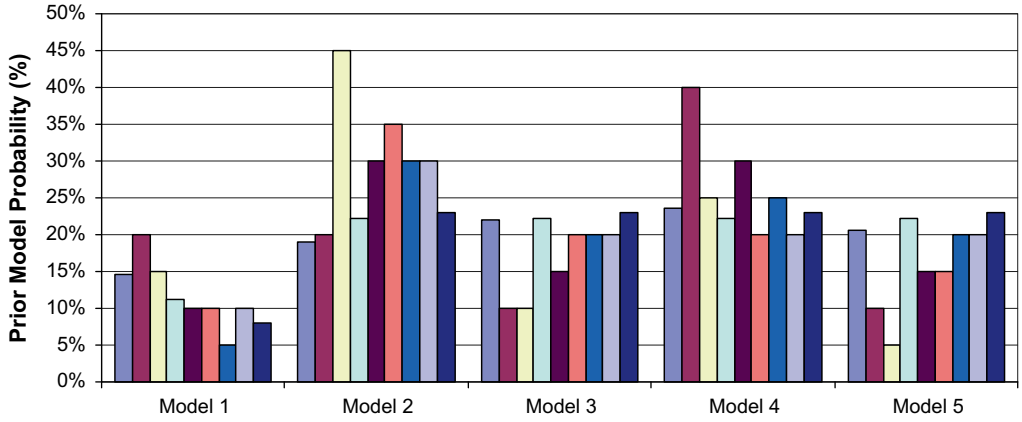

Figure 7 Column chart of the prior probability of the five geological models. The columns of each model represent elicited prior model probability from nine experts (from Pohlmann et al., 2007).

used experts with in-depth experience regarding mapping and geologic interpretations of the region. Fig. 7 shows the elicited prior model probability of the five models from nine experts, and Fig. 8 plots the aggregated prior probabilities using the iteration method (the aggregation results are again almost identical to those of a simple average). The aggregated prior model probabilities are also uniform. The uniform prior probabilities may be related to the inherent uncertainty embedded in the models, as well as the aggregation method (we are not aware of a better aggregation method). In addition, since the elicitation is for model probabilities, many techniques developed for model parameter probabilities (e.g., Cooke, 1991; O'Hagan, 1998; Bedford et al., 2006) are not directly applicable. The neutral choice of treating all models equally likely appears a reasonable selection of prior model probabilities in BMA.

Although the aggregated prior model probabilities are used in BMA to calculate the posterior model probabilities, it is also valuable to investigate the prior probabilities generated by each expert. It is likely that one expert gives a better evaluation than other experts, although his/her elicited probabilities may be different from the aggregated probabilities. For example, Expert 2 ranked the NIM1 model as the least plausible model, while the aggregated results show that the model is the most plausible one. However, without a rigorous (and time consuming) analysis (e.g., the cross-validation of Ye et al., 2005), it is unknown whether his/her estimation of the prior model probabilities is better. We propose a minimum entropy method to select the elicited probabilities from a single expert. The Shannon's entropy

$$H = -\sum_{k=1}^{K} p_k \log p_k \tag{9}$$

is the combined prior uncertainty measure of the models. The entropy is the expected value of $-\log p_k$, a measure of prior uncertainty associated with model M_k. When there is no information for evaluating the models and all models are treated equally likely ($p_k \equiv 1/K$), the entropy is the maximum ($H = \log K$). When more information is available to evaluate the models to reduce the model uncertainty, the entropy decreases. The smallest value it can attain corresponds to perfect certainty on the part of the analyst, *a priori*, that model M_k associated with some k would prove to be correct so that $p_k = 1$ and, by virtue of $\sum_{k=1}^{K} p(M_k) = 1$, $H = 0$. In this sense, when experts give different sets of prior model probabilities, the set of minimum entropy indicates the largest amount of information and the least amount of uncertainty. In this analysis, the prior probabilities of Expert 5 have the smallest entropy of 1.3, and the probabilities of the five models are 15%, 45%, 5%, 5% and 30%. These prior probabilities differ from the aggregated prior probabilities (25%, 31%, 11%, 13% and 20% for the five models), because the former set assigns larger confidence on models NIM1 and CMB2 but less on MME, NIM2 and CMB1. A potential difficulty with this minimum entropy approach is the lack of a guarantee that it would lead to optimum predictive performance.

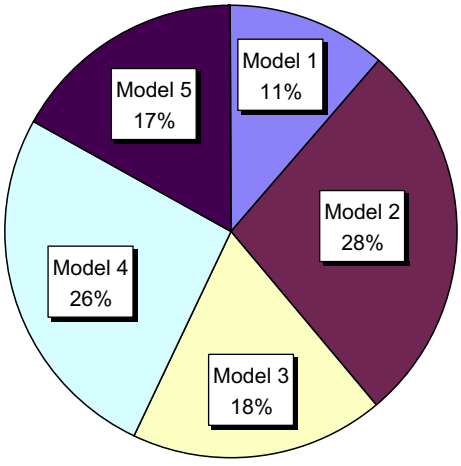

Figure 8 Iteratively aggregated prior probabilities of the five geological models (from Pohlmann et al., 2007).

When enough data are available, we prefer to use a cross-validation method to select the best set of prior model probabilities that gives the best predictive performance (Ye et al., 2005). The cross-validation separates the data set, **D**, into two parts, calibration part $\mathbf{D}^A$ and test part $\mathbf{D}^B$; $\mathbf{D}^A$ is used for calibrating the model to obtain the model

likelihood $p(\mathbf{D}^A|M_k)$ and subsequently the posterior model probability $p(M_k|\mathbf{D}^A)$ (Eq. (2)). Note that different sets of the prior model probabilities give different sets of the posterior model probabilities. The best set of prior model probabilities is selected by statistics of predicting the test data $\mathbf{D}^B$. Ye et al. (2005) proposed three statistics (i.e., log score, mean square prediction error and mean absolute prediction error), and the mean square prediction error (MSPE) is taken an example here,

$$\mathrm{MSPE} = \frac{1}{N^B} \sum_{d \in \mathbf{D}^B} \sum_{k=1}^{K} [(\hat{d}|M_k, \mathbf{D}^A) - d]^2 p(M_k|\mathbf{D}^A) \tag{10}$$

where N^B is the number of cross-validation data and $\hat{d}$ is a prediction of model M_k corresponding to the cross-validation data $d \in \mathbf{D}^B$. The best set of prior model probabilities corresponds to the smallest MSPE value. The cross-validation method can guarantee that the selected set of prior model probabilities yields the best prediction performance of BMA.

Conclusions

Five alternative recharge models have been independently developed for the DVRFS based on three different 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 NIM1 and NIM2 models distinguish each other by opposite (with and without) conceptualizations of the runon—runoff component. The CMB1 model has only the alluvial mask (below which recharge is zero), while the CMB2 model have the both alluvial and elevation masks. It is uncertain which model (or models) should be used for groundwater flow and contaminant transport modeling at the DVRFS, where recharge is a major driving force of groundwater flow. When the BMA method is used for assessing model uncertainty, it is essential to estimate the prior model probabilities. In addition, using informative prior probabilities can improve model predictive performance.

Expert elicitation is used in this study for evaluating, selecting and weighting the recharge models and the most important result of the elicitation is the prior model probability. The elicitation was conducted following the process of Keeney and von Winterfeldt (1991). The entire process took about 2 months using state and national experts with experience in the southwestern US. The experts were selected for their familiarity with recharge and hydrogeological conditions of the DVRFS and experience at the forefront of research in recharge estimation. The most important result of the elicitation is the prior model probability, $p(M_k)$, used to calculate the posterior model probabilities $p(M_k|\mathbf{D})$ for the BMA. Using the expert elicitation enables us to evaluate the models on the basis of expert judgment. In contrast to the common practice of treating alternative models equally likely and thus ignoring the prior information, using the informative prior model probability can improve the correspondence of the model predictions and measurements in BMA (Madigan et al., 1995; Zio and Apostolakis, 1996; Ye et al., 2005). Quantification of the improvement however is beyond scope of this paper.

Elicited prior model probabilities were aggregated using the simple averaging and iterative methods. Although the aggregation results are almost identical, the iterative method is preferred, since it considers between-expert variability. The NIM1 model received the largest prior probability, indicating that, on average, the experts had more confidence in this model. The confidence results from its comprehensive incorporation of the processes controlling net infiltration and potential recharge. However, one expert suspected the reliability of input data to this complicated model. The MME model on average received the second largest prior probability, with experts citing two reasons: that the model is simple and that it has been widely used in Nevada. It appears that the principle of parsimony was used by the experts to evaluate the alternative models. This is consistent with our intuition to try a simple model first and only move to a more complicated one when the simple one is inadequate. Whereas the principle of parsimony was not the only rule used by the experts, since the most complicated NIM2 model received the largest prior probability. The aggregated prior model probabilities do not support selecting a single recharge model and discarding others, suggesting that several models should be used in line with Scanlon et al. (2002).

Although each expert gave significantly different prior probabilities to the different models, the aggregated prior probabilities are close to the neutral choice that treats the five models equally likely and assigns 20% probability to each of the five models. The largest deviation is 11% for the NIM1 model. A similar phenomenon was also observed for an elicitation of five alternative geologic models (Pohlmann et al., 2007). This indicates the inherent uncertainty of the five recharge and geologic models, and is also related to the aggregation method. If a set of prior probability from a single expert is of more interest, we suggest selecting the probability set with the minimum Shannon's entropy. The minimum entropy implies the smallest amount of uncertainty and the largest amount of information used to evaluate the models. However, a potential difficulty with this minimum entropy approach is the lack of a guarantee that it will lead to the best predictive performance. When enough data are available, we prefer to use a cross-validation method to select the best set of prior model probabilities that gives the best predictive performance. Visualizing the recharge estimates and using pattern recognition to evaluate the alternative models appears a promising method for assessing the recharge model uncertainty (Lin and Anderson, 2003).

Acknowledgements

This research was supported in part by the US Department of Energy, National Nuclear Security Administration Nevada Site Office under Contract DE-AC52-00NV13609 with the Desert Research Institute. The first author conducted part of the research when he was employed by the Desert Research Institute. The authors are thankful to Randy Laczniak, Glendon Gee, Chuck Russell, Joe Hevesi and Greg Pohll for their participation in the elicitation.

Appendix A. Questionnaire used for the recharge model elicitation

Part I: Taking into consideration of the project, answer questions below for each recharge model.

1. To what degree is the model based on solid physical principles? (high, intermediate or low)
2. To what degree are the model assumptions solid and reasonable? (high, intermediate or low)
3. Are the model parameters measurable outside the context of the model? (yes, no)
4. What is the degree of sensitivity of the model outputs to model parameters? (high, intermediate or low)
5. To what degree is the model amenable to confirmation/validation on the basis of available measurements? (high, intermediate or low)
6. To what degree does model calibration demonstrate model plausibility? (high, intermediate or low)
7. To what degree may the model capture plausible future phenomena and events against which it cannot be presently assessed or calibrated? (high, intermediate or low)
8. To what degree does the model (concept, assumptions, implementation and results) agree with your knowledge and experience? (high, intermediate or low)
9. Is the model contrary to any of your knowledge and experience? (yes, no) If your answer is ''yes'', please specify the reason.
10. Is the model qualitatively comparable with others in terms of their plausibility? (yes, no) If your answer is ''no'', please specify the reason.

Part II: Taking into consideration the project, answer the questions below with your best estimates expressed as a point value.

11. Is the model set complete? (yes, no) If your answer is no, specify the additional plausible recharge model(s)?
12. Which model do you believe gives the best predictions of recharge?
13. What probability range (e.g., 40–60%) reflects your degree of belief that the model is the best?
14. Which model do you believe gives the worst predictions of recharge?
15. What probability range reflects your degree of belief that the model is the worst?

Part III: Taking into consideration the project, answer the questions below with your best estimates expressed as a point value.

16. What are the model ranks in terms of model plausibility? Models are ranked from 1 (the least plausible) to 5 (the most plausible). Different models may have the same rank, indicating that the expert has the same degree of belief as to the plausibility of the models.
17. What is the probability value that best represents the confidence you would place on each recharge model, given the objective of the analysis? Different models may have the same probability, indicating that the expert has the same degree of belief as to the plausibility of the models.

References

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

Apostolakis, G., 1990. The concept of probability in safety assessment of technological systems. Science 250, 1359–1364.

Arnell, N.W., Tompkins, E.L., Adger, A.N., 2005. Eliciting information from experts on the likelihood of rapid climate change. Risk Analysis 25 (6), 1419–1431.

Ayyub, B.M., 2001. Elicitation of Expert Opinions for Uncertainty and Risks. CRC Press, Boca Raton.

Bedford, T., Quigley, J., Walls, L., 2006. Expert elicitation for reliable system design. Statistical Science 21 (4), 428–450.

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.

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

Bonano, E.J., Apostolakis, G.E., 1991. Thoretical foundation and practical issues for using expert judgments in uncertainty analysis of high-level radioactive waste disposal. Radioactive Waste Management and the Nuclear Fuel Cycle 16 (2), 137–159.

Bonano, E.J., Hora, S.C., Keeney, R.L., von Winterfeldt, D., 1990. Elicitation and use of expert judgment in performance assessment for high-level radioactive water repository. Nuclear Regulatory Commission, NUREG/CR-5411, Washington, DC.

Bond, C.E., Gibbs, A.D., Shipton, Z.K., Jones, S., 2007. What do you think this is? ''Conceptual uncertainty'' in geoscience interpretation. GSA Today 17 (11), 4–10.

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multiple Model Inference: A Practical Information-theoretical Approach, second ed. Springer, New York.

Burnham, K.P., Anderson, D.R., 2004. Multimodel inference – understanding AIC and BIC in model selection. Sociological Methods & Research 33 (2), 261–304.

Carroll, R.W.H., Pohll, G.M., Earman, S., Hershey, R.L., 2007. Global optimization of a deuterium calibrated discrete-state compartment model (DSCM): application to the eastern Nevada Test Site. Journal of Hydrology 345 (3–4), 237–253.

Cooke, R.M., 1991. Expert in Uncertainty: Opinion and Subjective Probability in Science. Oxford University Press, New York.

Curtis, A., Wood, R. (Eds.), 2004. Geological Prior Information: Informing Science and Engineering. Geological Society of London, Special Publication, p. 239.

D'Agnese, F.A., Faunt, C.C., Hill, M.C., Turner, A.K., 1997. Hydrogeologic evaluation and numerical simulation of the Death Valley regional groundwater system, Nevada and California. US Geologic Survey Water Resources Investigation Report 96-4300.

Daly, C., Neilson, R.P., Phillips, D.L., 1994. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. Journal of Applied Meteorology 33, 140–158.

De Groot, M., 1974. Reaching a consensus. Journal of American Statistical Association 69, 118–121.

Dettinger, M.D., 1989. Reconnaissance estimates of natural recharge to desert basins in Nevada, USA, by using chloride-balance calculations. Journal of Hydrology 106 (1–2), 55–78.

DeWispelare, A.R., Herren, L.T., Clemen, R.T., 1995. The use of probability elicitation in the high-level nuclear waste regulation program. International Journal of Forecasting 11, 5–24.

Draper, D., 1995. Assessment and propagation of model uncertainty. Journal of Royal Statistical Society B 57 (1), 45–97.

Draper, D., Pereira, A., Prado, P., Saltelli, A., Cheal, R., Eguilior, S., Mendes, B., Tarantola, S., 1999. Scenario and parametric uncertainty in GESAMAC: a methoological study in nuclear waste disposal risk assessment. Computer Physics Communications 117, 142–155.

Epstein, B., 2004. Development and uncertainty analysis of empirical recharge prediction models for Nevada's Desert Basins. Masters Thesis, University of Nevada, 202 pp.

Feeney, T.A., Campana, M.E., Jacobson, R.J., 1987. A deuterium-calibrated groundwater flow model of the west Nevada Test Site and vicinity. DOE/NV/10384-16, Desert Research Institute, NV.

Garthwaite, P.H., Kadane, J.B., O'Hagan, A., 2005. Statistical methods for eliciting probability distributions. Journal of the American Statistical Association 100 (470), 680–700.

Geomatrix Consultants, 1998. Saturated zone flow and transport expert elicitation project. Deliverable Number SL5X4AM3. CRWMS M&O, Las Vegas, NV.

Ghabayen, S.M.S., McKee, M., Kemblowski, M., 2006. Ionic and isotopic ratio for identification of salinity sources and missing data in the Gaza aquifer. Journal of Hydrology 318 (1–4), 360–373.

Hardman, G., 1936. Nevada precipitation and acreages of land by rainfall zones. Nevada University Experiment Station Report and Map.

Harrill, J.R., 1976. Pumping and groundwater storage depletion in Las Vegas Valley, Nevada, 1955–1974. Nevada Division of Water Resources Bulletin 44, p. 70.

Hevesi, J.A., Flint, A.L., Flint, L.E., 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, US Geological Survey, Sacramento, CA.

Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T., 1999. Bayesian model averaging: a tutorial. Statistical Science 14 (4), 382–417.

Hora, S.C., Iman, R.L., 1989. Expert opinion in risk analysis: the NUREG-1150 methodology. Nuclear Science and Engineering 102, 323–331.

Hora, S., Jensen, M., 2005. Expert panel elicitation of seismicity following glaciation in Sweden. SSI Report 2005:20, Swedish Radiation Protection Authority.

Hora, S.C., von Winterfeldt, D., 1997. Nuclear waste and future societies: a look into the deep future. Technological Forecasting and Social Change 56, 155–170.

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

Jeffreys, H., 1957. Scientific Inference, seconnd ed. Cambridge University Press, Cambridge, UK.

Kashyap, R.L., 1982. Optimal choice of AR and MA parts in autoregressive moving average models. IEEE Transactions on Pattern Analysis and Machine Intellegience 4 (2), 99–104.

Keeney, R.L., von Winterfeldt, D., 1991. Eliciting probabilities from experts in complex technical problems. IEEE Transactions on Engineering Management 38 (3), 191–201.

Kirk, S.T., Campana, M.E., 1990. A deuterium-calibrated groundwater flow model of a regional carbonate-alluvial system. Journal of Hydrology 119, 357–388.

Lele, S.R., Das, A., 2000. Elicited data and incorporation of expert opinion for statistical inference in spatial studies. Mathematical Geology 32 (4), 465–487.

Lin, Y.-F., Anderson, M.P., 2003. A digital procedure for ground water recharge and discharge pattern recognition and rate estimation. Ground Water 41 (3), 306–315.

Madigan, D., Gavrin, J., Raftery, A.E., 1995. Eliciting prior information to enhance the predictive performance of Bayesian graphical models. Communications in Statistics: Theory and Methods 24, 2271–2292.

Maurer, D.K., Berger, D.L., Prudic, D.E., 1996. Subsurface flow to Eagle Valley from Vicee, Ash, and Kings Canyons, Carson City, Nevada, estimated from Darcy's Law and the chloride-balance method. US Geological Survey Water-Resources Investigation Report 96-4088, 38 pp.

Maxey, G.B., Eakin, T.E., 1949. Ground water in White River Valley, White Pine, Nye, and Lincoln Counties, Nevada. Nevada State Engineer, Water Resource Bulletin, No. 8 (prepared in cooperation with the United State Department of the Interior Geological Survey), Carson City, Nevada.

McKenna, S.A., Walker, D.D., Arnold, B., 2003. Modeling dispersion in three-dimensional heterogeneous fractured media at Yucca Mountain. Journal of Contaminant Hydrology 62 (3), 577–594.

Miklas, M.P.J., Norwine, J., DeWispelare, A.R., Herren, L.T., Clemen, R.T., 1995. Future climate at Yucca Mountain, Nevada proposed high-level radioactive waste repository. Global Environmental Change 5 (3), 221–234.

Minor, T.B., Russell, C.E., Mizell, S.A., 2007. Development of a GIS-based model for extrapolating mesoscale groundwater recharge estimates using integrated geospatial data sets. Hydrogeology Journal 15 (1), 183–195.

Morgan, M.G., Keith, D.W., 1995. Subjective judgments by climate experts. Environmental Policy Analysis 29 (10), 468–476.

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.

O'Hagan, A., 1998. Eliciting expert beliefs in substantial practical applications. The Statistician 47 (1), 21–35.

O'Hagan, A., Oakley, J.E., 2004. Probability is perfect, but we can't elicit it perfectly. Reliability Engineering & System Safety 85, 239–248.

Parent, E., Bernier, J., 2003. Encoding prior experts judgments to improve risk analysis of extreme hydrological events via POT modeling. Journal of Hydrology 283, 1–18.

Pike, W.A., 2004. Modeling drinking water quality violations with Bayesian networks. Journal of the American Water Resources Association 40 (6), 1563–1578.

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

Pohlmann, K., Ye, M., Reeves, D., Zavarin, M., Decker, D., Chapman, J., 2007, Modeling of groundwater flow and radionuclide transport at the climax mine sub-CAU, Nevada Test Site, DOE/NV/26383-06. Nevada Site Office, National Nuclear Security Administration, US Department of Energy, Las Vegas, NV.

Refsgaard, J.C., van der Sluijs, J.P., Brown, J., van der Keur, P., 2006. A framework for dealing with uncertainty due to model structure error. Advances in Water Resources 29, 1586–1597.

Rehfeldt, K. (Ed.), 2004. Hydrologic Data for the Groundwater Flow and Contaminant Transport Model of Corrective Action Units 101 and 102: Central and Western Pahute Mesa, Nye County, Nevada, Stoller-Navarro Joint Venture, Las Vegas, NV.

Russell, C.E., 2004. Documentation of data and method for EDCMB extended analysis. Desert Research Institute, Division of Hydrologic Sciences Letter Report.

Russell, C.E., Minor, T., 2002. Reconnaissance estimates of recharge based on an elevation-dependent chloride mass-balance approach, DOE/NV/11508-37, Publication No. 45164. Prepared for the US Department of Energy, National Nuclear Security Administration Nevada Operations Office. Desert Research Institute, Las Vegas, NV.

Sadler, W.R., Campana, M.E., Jacobson, R.J., Ingraham, N.L., 1991. A deuterium-calibrated, discrete-state compartment model of regional groundwater flow, Nevada Test Site and vicinity, DOE/NV/10845-09, Publication Number #45088. Desert Research Institute.

Scanlon, B.R., 2004. Evaluation of methods of estimating recharge in semiarid and arid regions in the southwestern US. In: Hogan, J.F., Philips, F.M., Scanlon, B.R. (Eds.), Groundwater Recharge in a Desert Environment: The Southwestern United States. American Geophysical Union, pp. 235–254.

Scanlon, B.R., Healy, R.W., Cook, P.G., 2002. Choosing appropriate techniques for quantifying groundwater recharge. Hydrogeology Journal 10, 18–39.

Schwarz, G., 1978. Estimating the dimension of a model. Annual Statistics 6 (2), 461–464.

Shannon, C.E., 1948. A mathematical theory of communication. Bell System Technology Journal 27, 379–423, 623–656.

Stiber, N.A., Pantazidou, M., Small, M.J., 1999. Expert system methodology for evaluation reductive dechlorination at TCE sites. Environmental Science and Technology 33 (17), 3012–3020.

Stiber, N.A., Small, M.J., Pantazidou, M., 2004. Site-specific updating and aggregation of Bayesian belief network models for multiple experts. Risk Analysis 24 (6), 1529–1538.

Vrugt, J.A., Robinson, B.A., 2007. Treatment of uncertainty using ensemble methods: comparison of sequential data assimilation and Bayesian model averaging. Water Resources Research 43 (1), W01411. doi:10.1029/2005WR004838.

Vrugt, J.A., Clark, M.P., Diks, C.G.H., Duan, Q., Robinson, B.A., 2006. Multi-objective calibration of forecast ensembles using Bayesian model averaging. Geophysical Research Letter 33, L19817. doi:10.1029/2006GL027126.

Wagener, T., Gupta, H.V., 2005. Model identification for hydrological forecasting under uncertainty. Stochastic Environmental Research and Risk Assessment 19, 378–387.

Wingle, W.L., Poeter, E.P., 1993. Uncertainty associated with semivariograms used for site simulation. Ground Water 31 (5), 725–734.

Ye, M., Neuman, S.P., 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., Neuman, S.P., Meyer, P.D., Pohlmann, K., 2005. Sensitivity analysis and assessment of prior model probabilities in MLBMA with application to unsaturated fractured tuff. Water Resources Research 41 (12), W12429. doi:10.1029/2005WR004260.

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

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