Ranking of conceptualized groundwater models based on model information criteria

Majid Ehtiat, S. Jamshid Mousavi and Abbas Ghaheri

ABSTRACT

Modelers typically test alternative conceptualizations of a groundwater system because the ‘true’ hydrogeological model of the system is unknown, and the observed data are limited. This study elaborates on how to rank different conceptualizations of a groundwater system by using selection techniques and criteria that make a balance between complexity and accuracy of models. Depending on the parameterization method, number of spatial zones in parameterization, and the calibration approach, 30 different models have been built and compared against 12 discriminating selection criteria in the Dehloran groundwater system application in Iran. The validation results have shown that the best model calibrated by a two-step approach is far superior than the one calibrated by a one-step approach. Nevertheless, the low ranked models of the first step should not be disregarded, because the high ranked models selected in the first and second steps of calibration are not necessarily the same. In addition, although the ratio of the number of total observations to the number of parameters strongly affects the performance of a selection criterion, a model’s complexity assessment term of a number of selection criteria is not continuous with respect to the number of parameters that causes an inappropriate performance of these criteria at some certain points.

Key words | conceptualization, groundwater modeling, multi-modeling, selection criteria

INTRODUCTION

In recent years, researchers have actively sought to construct aquifer simulation models as a primary tool for groundwater resource management. Groundwater modeling invariably involves simplifying a real system into a conceptual model to be used for analysis purposes. Such a conceptual model must capture the essence of the problem (Bredehoef 2003). A number of researchers have emphasized that conceptualization is the most important and difficult step in groundwater modeling (Samper et al. 1990; Kolm & Van Der Heijde 1996; Poeter & Hill 1997; Bredehoef 2003, 2005). Recent studies have clearly shown that a large degree of uncertainty arising in parameter estimation, predictions and results of a groundwater model lies in conceptual model deficiencies (e.g., Hyun & Lee 1998; Bredehoef 2003, 2005; Sonnenborg et al. 2003; Poeter & Anderson 2005; Gaganis & Smith 2006; Rojas et al. 2010).

Modelers are faced with an enormous volume of data in practical cases. As well, in some cases the available field data have many shortcomings. In the first case, more than one description of the system may result from the conceptualization step (Samper et al. 1990). In the latter case, in order to ensure that the true representation of the real system is reflected in the model construction, one should develop and test a set of candidate models. The candidate models typically differ in the way that they conceptualize the real system.

Given the uncertainties in modeling (conceptualization, parameterization, and calibration), a groundwater model can only be used to approximate the behavior of a subsurface system. Therefore, one can merely arrive at the best approximation (‘best’ model) and/or make inference from a set of ‘good’ models. A good model is an effective one.
that has as much transparency (understandability of the model dynamics) as accuracy (ability to simulate the behavior of the real system). Perhaps, we cannot do any better than to use our hydrogeologic insight to develop what we believe is a full set that includes the true situation, and then admit that we are merely making our best hydrogeologically informed guess (Voss 2012b).

The above-mentioned recognition has led to a growing tendency among hydrologists to postulate several alternative hydrologic models for a site by using a multi-modeling approach and use various criteria to (1) rank these models, (2) eliminate some of them, and/or (3) weight and average predictions and statistics generated by multiple models (Ye et al. 2008). Model selection techniques have been applied in several groundwater modeling studies (Hyun & Lee 1998; Poeter & Anderson 2005; Foglia et al. 2007; Tan et al. 2008; Rojas et al. 2010; Ye et al. 2010; Engelhardt et al. 2012).

In this study, some new model selection techniques have been used, and their performances have been evaluated. The paper particularly focuses on how to rank different conceptualizations or models while using different selection criteria. In this regard, the steps of a model selection process are first reviewed, and the by-product products that can be obtained from model selection criteria are then introduced. To this end, several criteria are introduced and examined as effective, yet practical approaches for ranking candidate models and, consequently, selecting the best one or defining a set of good models. Finally, model selection techniques and criteria are evaluated within a multi-modeling approach in the Dehloran aquifer system located in an arid western part of Iran, and some practical points related to the appropriate use of selection criteria are provided.

**MODELING TOOLS AND METHODOLOGY**

**Model selection criteria**

While evaluating several alternative groundwater simulation models based on different conceptualizations, we need to calibrate each of the models. To do so, automatic inverse modeling is often used as a state-of-the-art convenience to estimate hydrogeological parameters of interest. Then, model selection criteria are used for assessing and ranking the candidate models as well as making some beneficial inferences. In this study, we use a number of selection criteria whose theoretical principles and formulas are discussed below.

It has been generally understood that as parameters are added to a model, the modeling error decreases, and the fit improves although the model uncertainty increases, and the predictions become less accurate (Yeh 1986; Poeter & Anderson 2005; Hill 2006). Thus, comparisons are to be made among several models in terms of both the models’ complexity and the modeling error indisputably. Model selection criteria or information criteria (IC) are used for this purpose.

Table 1 elaborates on the formulas and elements of 12 model selection criteria being investigated in this study. All the formulas comprise a term representing lack of fit (L) and a bias correction factor quantifying a model's complexity. With regard to lack of fit, there is usually a need to calculate the maximum likelihood estimate of the residual variance, $\sigma^2$. For this case, $\log(L) = -(N/2) \log(\sigma^2)$ where $\sigma^2 = (WRSS/N)$ (Poeter & Anderson 2005; Singh et al. 2010). WRSS is the weighted residual sum of squared, and N is the number of total observations equal to the number of observations, NO, plus the number of prior estimation equations, NP (Hill & Tiedeman 2007). Thus, the term $-2 \log(L)$ can be replaced with $N \log \left( \frac{WRSS}{N} \right)$ in Table 1.

The second term penalizes models with greater complexity and is a function of $K$ and $N$ where $K$ is the number of estimated parameters. In the case of least-squares regression, the value of $K$ must be increased by 1 to reflect the variance estimate ($\sigma^2$) as an extra parameter (Poeter & Anderson 2005; Symonds & Moussalli 2011). The penalty term measures the increased unreliability of the chosen model due to the increased number of parameters (Mutua 1994).

Figure 1 illustrates the penalty term values of different criteria for $N = 5, 10, 20, 50$ and $K = 1, \ldots, 100$. One can see that the penalty term in AIC and KIC is constant for a given $K$. For CAIC, SIC, DIC, and HIC, increasing $N$ with a constant $K$ enlarges the value of penalty term significantly, whereas a contrary behavior is observable for AICu, RICc, and KICc. This figure also shows that the penalty terms of AICc, SICc, and HICc are treated differently compared to other criteria.
In general, the variations of the penalty terms along with changes in $K$ are low for a given $N$. However, when $K$ is equal to $N - 1$, the penalty term cannot be calculated for AICc. This condition occurs for SICc and HICc as well when $K$ is equal to $N - 2$. Similarly, the penalty terms are incomputable for AICu and KICc when $K \geq N - 1$ (except in point $K = N$ for KICc) and for RICc when $K \geq N - 2$ (except in point $K = N - 1$). The penalty values of AICc, SICc, and HICc are relatively high and vary sharply when $K$ is located near $N$. Furthermore, significant changes in value and sign can occur for the penalty term as a result of minor variations in $K$. For example, when $N = 50$, for $K = 46, 48,$ and $50$, the values of AICc’s penalty term will be $1,533, 4,800, -5,000$, respectively. This indicates that AICc may behave irrationally around the point at which $K = N - 1$ with two different asymptotically converging limits of the penalty values from left and right sides (see Figure 1 at $K = N - 1$). This situation proves true for AICu, (KICc) and RICc when $K$ is selected approximately near $N - 1$ and $N - 2$. This indicates that all these criteria (AICc, SICc, HICc, HICu, KICc, and RICc) may behave irrationally around a certain point ($K = N$ or $N - 1$ or $N - 2$, depending on the criterion) at which the penalty term is not continuous and would converge asymptotically to different limits from different sides (see Figure 1).

The comparison of SIC with AIC and AIC3 reveals that the SIC’s penalty term is larger than that of AIC and AIC3 when $N$ is greater than $10^{2}$ and $10^{3}$, respectively. In addition, it is evident that RICc has a larger penalty function compared to AIC and a smaller penalty compared to AICc when $K \gg K$ (Leng 2015). If $K$ is considered as the primary measure of complexity (Hill 2006; Voss 2011b), unlike AICu, RICc, and KICc,
the penalty terms of CAIC, SIC, DIC, and HIC will increase by adding the observed data when the complexity of the model is kept constant. Such a result may not be justified as it is expected that the penalty term (model complexity) increases while increasing $N/K$.

AIC has a tendency to overfit when the sample size $N$ is small or when $K$ is a moderate-to-large fraction of the sample size $N$ (Hurvich & Tsai 1989). Compared to AIC, AICc has an additional nonstochastic penalty term that is for small-sample bias adjustments (Hurvich & Tsai 1989) and, consequently, tends to select much better models. AICc has been recommended where $N/K$ is less than 40 (Burnham & Anderson 2002; Hill & Tiedeman 2007). When $N$ becomes larger than $K$, the additive coefficient converges to 1, and AICc approaches AIC. Compared to AICc, AICu has an additional penalty term that provides better model choices than AICc for moderate-to-large sample sizes $N$, except when the ‘true’ model is of infinite order (Rao & Wu 2001). Contrary to the AIC, CAIC practically never overfits, but it often underfits (Dziak et al. 2012).

SIC can overfit in small samples the same way as AIC does (McQuarrie 1999). The models selected by SIC tend to contain fewer parameters than those selected by AIC. In the same vein, SICc shows a better performance than SIC in small samples.
The penalty term in DIC is typically similar to that of SIC when the number of observations \( N \) is replaced with \( N/2\pi \). As argued by Draper (1995), this correction is necessary in order to improve the performance in finite samples (Figure 1). As well, HIC is obtained by substituting the \( \log(N) \) factor of the penalty term in SIC with a slower diverging quantity, i.e., \( \log(\log(N)) \). HICc is a small-sample bias correction version of HIC. KIC (or AIC3) is a more penalized version of AIC that places more emphasis on parsimony than AIC does. KICc is known as a large-sample based selection criterion where under-fitting is not that much problematic, yet over-fitting still remains probable (Cavanaugh 2004).

The above explanations and analyses reveal that the behavior of different criteria and their orders of magnitude with respect to \( N, K \), and \( N/K \) may be complicated or even misleading. Comparative measures focusing on their
relative values could, therefore, be of more help to clarify their relative advantages and disadvantages.

**Assessment measures**

**Δ value**

Given the aforementioned IC, the model with the lowest value, IC\textsubscript{min}, may be chosen as the ‘best’ model, and the remaining models are ranked from best to worst. High ranked models constitute a set of ‘good’ models. In addition, IC differences (Δ values), \( \Delta IC_i = IC_i - IC_{\text{min}} \), where \( i = 1, 2, \ldots, R \), can be used in order to rank the competing models (Burnham & Anderson 2004). Δ value for each model is a representation of the estimated distance to the best model.

**Evidence ratio**

To quantify the plausibility of each model, one needs an estimate of the likelihood of the model to be the best
approximation for a given data set, $\mathcal{L}$ (model | data). This term is equal to $\exp(-0.5\Delta IC_i)$ (Akaike 1981; Burnham & Anderson 2004). The ratio of likelihood of model $i$ to that of model $j$ is termed as the evidence ratio (ER):

$$ER = \frac{\exp(-0.5\Delta IC_i)}{\exp(-0.5\Delta IC_j)}$$

This quantity provides a measure of the extent to which model $i$ is more likely to be the best model than model $j$, especially when $i$ is the best model (Poeter & Anderson 2005).

Model weight

A posterior model weight, $w_i$, is obtained for model $i$ by normalizing the model’s likelihood in such a way that the denominator is simply the sum of the relative likelihoods for all candidate models:

$$w_i = \frac{\exp(-0.5\Delta IC_i)}{\sum_{j=1}^{R} \exp(-0.5\Delta IC_j)}$$

In this way, two models with the same $\Delta IC$ are given the same weight regardless of the values of their penalty terms, where $w_i$ can be interpreted as the weight of evidence or the probability that model $i$ is the best model, so $\sum w_i = 1$. Unlike the ER, $w_i$ values depend on the full set of models. A model with the greatest $w$ has the lowest value of $\Delta$. In the case of the AIC family, these weights or model probabilities are formal probabilities, yet they differ from Bayesian posterior probabilities (Lukacs et al. 2010).

Using the model weights, one can compare specific subsets of models with other subsets in the same way (e.g., all models containing an interaction term versus those that do not) (Wagenmakers & Farrell 2004; Lavoue & Droz 2009). For the sake of illustration, consider the candidate models sorted in two classes of A and B containing two types of different parameters (e.g., hydraulic conductivity (HK) and specific storage). Without loss of generality, assume that

$$\left(\sum_{i=1}^{j} w_{A_i}\right) / \left(\sum_{j=1}^{R} w_{B_j}\right) = q \geq 1.$$ 

Then, models of class A are $q$ times more likely than those of class B. The relative importance of individual parameters can also be investigated using weights, $w_i$. To this end, the weights for each model containing the parameters of interest are summed, which estimate the relative importance of the parameters in question for the available data and candidate models (Symonds & Moussalli 2011).

Rejection rule

Some simple rules of thumb have been introduced as cut-off rules to eliminate uninformative models. As discussed by Burnham & Anderson (2004), Poeter & Anderson (2005), and Burnham et al. (2011), the models with $\Delta \leq 2$ have substantial support and are regarded as very good models, and those with $2 < \Delta \leq 7$ have considerably less empirical support and should rather be dismissed. Finally, models with $\Delta > 10$ have essentially no support and can be definitely dismissed from further consideration. Richards (2005, 2008) explicated a cut-off border whereby models with $\Delta \leq 6$ should be considered for making inferences, and the rest are disregarded. However, laying some models away using a defined cut-off threshold for $\Delta$ values seems not to be a good way for multi-model based inferences because: (1) sometimes models with lower $\Delta$ values have relatively low weights, $w$ (say <0.5), meaning that the retained models as well as the ones deleted from the candidate set have less chance to be regarded as ‘good’ models; and (2) for $\Delta$ values gathered around the cut-off borders, there is no explicit distinction between desirability of models located up or down. These $\Delta$ values are on a continuous scale of information and are interpretable regardless of the measurement scale and whether the data are continuous, discrete, or categorical (Burnham et al. 2011).

A rather less Procrustean approach is to produce a ‘confidence set’ or ‘credibility set’ of models that are most realistically likely to be the best approximating model (Symonds & Moussalli 2011). This confidence set can be achieved by adding the weights, $w_i$, from the best downwards. If the sum of the highest $w_i$ values, $i = 1, \ldots, j; j \leq R$, is $\approx 95\%$ (or any other arbitrary value), one can say that the best model is within this subset by 95% confidence and could decide to base inferences only on these models. Another reasonable basis for constituting a...
confidence set of candidate models is selecting the models with weights that are within 10% of the highest (see e.g., Burnham & Anderson (2002) and Omole et al. (2015)).

GROUNDWATER MODEL SETUP

Study area

Dehloran regional groundwater flow system encompasses nearly 568.5 km² and is located in the west of Iran about 228 km distance from Ilam City (Figure 2). Dehloran plain lies between 32°22' and 32°45' northern latitude and 47°05' and 47°50' eastern longitude. Dehloran City is situated in the north of the plain, and the underlying groundwater resource serves a leading role in securing irrigation demand for the region. Two major agricultural zones are considered to be the primary consumers of the groundwater resource in the south-east and western outskirt of the plain. In addition, there is a municipal pumping well field in the south-east of the plain, above irrigation areas, which supplies the necessary drinking water for Dehloran City. The study area covers 360 km².

The analysis of results from geoelectrical profiles data (Mahab Ghods Consulting Engineers 1992) has led to the recognition of geologic layers, lithological cross sections, the aquifer geometry and the layers’ properties. The aquifer depth is within the limits of 140 m in the center of the plain, reaching 150 m in some points and decreases to 50 m and 5 m in the south and north, respectively. There are 22 observation wells, of which 18 wells are located in our modeling domain. According to data collected in 2007–2008 (Mahab Ghods Consulting Engineers 2008), there are a total of 233 operational wells in Dehloran well-field with about 67.0 MCM (million cubic meters) extraction of water annually, of which 115 wells with 38.2 MCM annual water extraction are located in the modeling domain.

Parameterization method

The zonation method (ZM) and the regularized pilot points (RPP) (Doherty 2003; Moor & Doherty 2006) method have been used to parameterize unknown parameters of the aquifer.

In the ZM, the model domain is divided into multiple subdomains or zones, and a constant value of a parameter of interest is assigned to each zone. These constants may then be considered as unknowns being adjusted in a calibration process (Prasad & Rastogi 2001; Rojas & Dassargues 2007).

The RPP method aims to provide a middle ground between cell-by-cell variability and the reduction to a few homogeneous zones for characterizing natural-world heterogeneity in groundwater models (Doherty & Hunt 2010). In this approach, a set of 2D scatter points (pilot points) are initially distributed throughout the model domain. The values assigned to the points of interest are then estimated by calibration. Afterward, the ‘true’ values for other cells of the model are calculated by using a spatial interpolation technique such as Kriging. In the RPP method (Alcolea et al. 2006), the initial values of pilot points must be in a geologically reasonable range expected in the study area. Regularization imposes an additional knowledge as well as a geological technical perspective, in the form of prior
information, on the parameter estimation process and can be integrated by the method of pilot points (Doherty 2005; Moor & Doherty 2006; Franssen et al. 2009; Doherty & Hunt 2010). In fact, the addition of each new parameter will be accompanied by at least one (possibly many) ‘regularization observations’, thus restoring the numerical predominance of observations over parameters (Doherty 2005).

HK and specific yield (Sy) are considered as unknown parameters to be determined by inverse modeling (see Table 2). A SWAT (Soil and Water Assessment Tool; Arnold et al. 1998) model of the investigated area has been developed and calibrated by Ehtiat et al. (2015) to estimate time-variant recharge rates to the aquifer. The ZM has been used to model spatial distribution of recharge and Sy, and both the methods of RPP and zonation have been used for modeling spatial variations of HK. We have also considered the zonation patterns of Sy the same as those of HK, as illustrated in Figure 5.

Conceptual models

Several alternative conceptualizations have been considered based on: (1) the parameterization methods of HK and Sy, (2) the number of spatial zones or pilot points in parameterization, and (3) the calibration approach (one-step or two-step) (Table 2). Accordingly, the developed models have been grouped into two types of A and B based on

<table>
<thead>
<tr>
<th>Group num.</th>
<th>Calibration state</th>
<th>Index</th>
<th>Unknown parameters</th>
<th>Parameterization method</th>
<th>Num. of parameters (K)</th>
<th>Num. of total observation (N)</th>
<th>N/K</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>Steady-state</td>
<td>A1</td>
<td>*</td>
<td>Zonation</td>
<td>2</td>
<td>18</td>
<td>9.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A2</td>
<td>*</td>
<td>Pilot point</td>
<td>4</td>
<td>18</td>
<td>4.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A3</td>
<td>*</td>
<td></td>
<td>8</td>
<td>18</td>
<td>2.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A4</td>
<td>*</td>
<td></td>
<td>13</td>
<td>18</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A5</td>
<td>*</td>
<td></td>
<td>19</td>
<td>18</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B1</td>
<td>*</td>
<td>*</td>
<td>21</td>
<td>38</td>
<td>1.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B2</td>
<td>*</td>
<td>*</td>
<td>31</td>
<td>48</td>
<td>1.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B3</td>
<td>*</td>
<td>*</td>
<td>46</td>
<td>63</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B4</td>
<td>*</td>
<td>*</td>
<td>55</td>
<td>72</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B5</td>
<td>*</td>
<td>*</td>
<td>76</td>
<td>93</td>
<td>1.2</td>
</tr>
<tr>
<td>Group 2</td>
<td>Transient</td>
<td>A6</td>
<td>*</td>
<td>*</td>
<td>2</td>
<td>216</td>
<td>108.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A7</td>
<td>*</td>
<td>*</td>
<td>4</td>
<td>216</td>
<td>54.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A8</td>
<td>*</td>
<td>*</td>
<td>8</td>
<td>216</td>
<td>27.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A9</td>
<td>*</td>
<td>*</td>
<td>13</td>
<td>216</td>
<td>16.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A10</td>
<td>*</td>
<td>*</td>
<td>19</td>
<td>216</td>
<td>11.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B6</td>
<td>*</td>
<td>*</td>
<td>2</td>
<td>216</td>
<td>108.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B7</td>
<td>*</td>
<td>*</td>
<td>4</td>
<td>216</td>
<td>54.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B8</td>
<td>*</td>
<td>*</td>
<td>8</td>
<td>216</td>
<td>27.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B9</td>
<td>*</td>
<td>*</td>
<td>13</td>
<td>216</td>
<td>16.6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B10</td>
<td>*</td>
<td>*</td>
<td>19</td>
<td>216</td>
<td>11.4</td>
</tr>
<tr>
<td>Group 3</td>
<td>Transient</td>
<td>A11</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>3</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A12</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>7</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A13</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>15</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A14</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>25</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td></td>
<td>A15</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>37</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B11</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>22</td>
<td>236</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B12</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>34</td>
<td>246</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B13</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>53</td>
<td>261</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B14</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>67</td>
<td>270</td>
</tr>
<tr>
<td></td>
<td></td>
<td>B15</td>
<td>*</td>
<td>*</td>
<td>*</td>
<td>94</td>
<td>291</td>
</tr>
</tbody>
</table>
parameterization methods used for unknown parameters. The ZM is used for parameterization of both HK and $S_y$ in the type A models, whereas in the type B models, the RPP and the ZMs are employed for HK and $S_y$ parameterizations, respectively.

Two calibration procedures have also been used to estimate the parameter values, namely two-step and one-step calibration approaches. In the two-step approach, for each type A or type B model, there are five sub-models (A1–A5 and B1–B5) for which a steady-state groundwater model is first calibrated. Then, the resulting spatial distributions of HK are used in the subsequent transient models (A6–A10 and B6–B10), where zonal $S_y$ values are calibrated. However, in the one-step approach, both HK and $S_y$ are calibrated concurrently by transient models A11–A15 and B11–B15.

Zonal patterns of $S_y$ for type B models are the same as those for the corresponding type A models with 1, 3, 7, 12, and 18 zones, respectively. The estimated distribution of HK for models B1–B5 and B11–B15 has been characterized using 20, 30, 45, 54, and 75 pilot points, respectively. In the
steady-state calibration mode, the differences among models A1–A5 is in the number of zones, and the number of pilot points and their distributions are different for models B1–B5. As a result, a total of 30 models have been developed for comparison purposes.

The US Geological Survey modular flow simulator MODFLOW (McDonald & Harbaugh 1988) and PEST optimizer (Doherty 2002) have been employed for simulation and optimization (automatic calibration) of all the models, respectively. Pilot points in combination with the regularization capabilities of PEST have been exploited in the way described by Doherty (2005). Assigning initial values and emplacement of pilot points have also been coordinated considering the attention points given by Doherty & Hunt (2010). Specifically, the pilot points have been located where results from a pumping test have been available. Therefore, some of the pilot point initial values have been set to estimated K values, and others have been assumed according to the K map of the study area. Lower and upper bounds of K have also been selected based on the values recommended in previous studies.

The steady-state and transient modes of calibration have been conducted based on the data on water levels from October 2005 to September 2006. The recharge rates have been calculated on a daily basis by SWAT, and the groundwater models have been calibrated using annual and monthly time steps in steady-state and transient modes, respectively.

To attain an acceptable simulated behavior of a subsurface system, the residuals between observed and simulated values must be limited. Acceptable residuals may vary for different hydraulic head calibration targets within different models (e.g., Henriksen et al. 2003; Rojas & Dassargues 2007). In this study, a calibration target has arbitrarily been defined as a root mean squared error (RMSE) lower than 1 m, which is about 1.0% of the difference between the highest and the lowest heads across the site.

**RESULTS AND DISCUSSION**

**Calibration results**

Figure 4 demonstrates the results of the steady-state and the transient modes of automatic calibration, including the RMSE values and the WRSS errors between simulated and observed water levels at the observation wells. In the case of steady-state calibration, RMSE values for the type B models with RPP method of parameterization are significantly less than those for the models of type A. Nevertheless, the calibration target is met only in model B5 and partly in model B4. In addition, the results indicate that the modeling error decreases for both types of models when the parameter dimension increases. This behavior is also observable in transient calibration of type B models while calibrating S_y. However, for the models of type A and in the same calibration mode, the maximum and minimum values of RMSE are associated with models A7 and A9, respectively. In transient calibration of S_y, the calibration target is met in models B8–B10 and partly in models A6 and A7.

The models of type A and type B calibrated by the one-step approach are found to outperform the corresponding models calibrated by the two-step approach. Nonetheless, the improvement for models of type A is significantly greater than that for type B models in this case. It shows the flexibility of the RPP method compared to the ZM. This is because the estimation of spatially distributed parameters by means of RPP often involves, to some extent, the imposition of smoothness of equality constraints of the inverse calibration problem, and therefore creates more confidence with respect to uniqueness and stability of the identified set of parameters. As well, an acceptable characterization of HK variations over the model domain is attainable by using the RPP method in the two-step calibration approach. In the one-step calibration approach, RMSE values of type B models decrease uniformly while increasing the number of pilot points, whereas in the case of the ZM, the largest RMSE belongs to model A13 and then decreases with the rise of the number of zones. In this mode, the calibration target is met in all the models except for model A13.

**Ranking of models**

To rank the models, models A1–A5 plus B1–B5, models A6–A10 plus B6–B10, and finally, models A11–A15 plus B11–B15 have been categorized into three separate groups of 1, 2, and 3, respectively, and then ΔIC_i and w_i values for each model i of every group have been calculated, as presented in Table 3.
The results indicate that AIC, CAIC, SIC, DIC, HIC, RICc, and AIC3 criteria select models B5, B10, and B15, respectively, as the ‘best’ model, although they are the most complex models (large number of parameters) among the models of each group. In other words, they select the transient model in group 2 in which the estimated parameters of HK have been obtained by the corresponding steady-state model in group 1. The selected models have been characterized with the RPP method and have shown the lowest RMSE. These selection criteria have opted for the aforementioned models with \( w = 1 \) among the candidate models of groups 1, 2, and 3 (note that just CAIC has assigned a weight equal to 0.999 to model B10). As a result, the rejection rule approach and the idea of making a confidence set have not practically been useful for these criteria in this particular case study.

Model A5 has been selected by AICc and SICc as the best model with \( w = 1 \) among the candidate models of group 1. It is worth noting that the values of penalty terms in these criteria are confusing. That is to say, the number of parameters for model A5 is near \( N \), causing a large negative value for the penalty term. For example, the penalty term values for candidate models in group 1 (models A1–A5 and B1–B5) using the AICc criterion are 4.8, 11.1,
<table>
<thead>
<tr>
<th>Group</th>
<th>Index</th>
<th>AICc</th>
<th>AICu</th>
<th>CAIC</th>
<th>SIC</th>
<th>SICc</th>
<th>DIC</th>
<th>HIC</th>
<th>HICc</th>
<th>RICc</th>
<th>AIC3</th>
<th>KICc</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>num.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>-8.1</td>
<td>170.4</td>
<td>-7.3</td>
<td>554.5</td>
<td>-5.8</td>
<td>0.0</td>
<td>0.655</td>
<td>7.5</td>
<td>97.5</td>
<td>0.0</td>
<td>95.5</td>
<td>0.8</td>
</tr>
<tr>
<td>A2</td>
<td>-6.1</td>
<td>172.5</td>
<td>-5.0</td>
<td>538.5</td>
<td>-0.5</td>
<td>5.3</td>
<td>0.655</td>
<td>5.1</td>
<td>99.8</td>
<td>0.0</td>
<td>95.5</td>
<td>0.8</td>
</tr>
<tr>
<td>A3</td>
<td>-9.9</td>
<td>177.5</td>
<td>-15.1</td>
<td>576.6</td>
<td>20.5</td>
<td>26.5</td>
<td>0.0</td>
<td>11.1</td>
<td>106.0</td>
<td>0.0</td>
<td>95.5</td>
<td>0.8</td>
</tr>
<tr>
<td>A4</td>
<td>6.4</td>
<td>184.8</td>
<td>9.479</td>
<td>459.9</td>
<td>10.4</td>
<td>9.2</td>
<td>0.5</td>
<td>11.1</td>
<td>114.0</td>
<td>0.0</td>
<td>95.5</td>
<td>0.8</td>
</tr>
<tr>
<td>A5</td>
<td>18.5</td>
<td>196.0</td>
<td>-36.3</td>
<td>3.0</td>
<td>1.000</td>
<td>NC</td>
<td>NC</td>
<td>25.3</td>
<td>128.2</td>
<td>0.0</td>
<td>95.5</td>
<td>0.8</td>
</tr>
<tr>
<td>B1</td>
<td>10.5</td>
<td>180.0</td>
<td>4.73</td>
<td>408.8</td>
<td>0.615</td>
<td>6.7</td>
<td>0.0</td>
<td>1.7</td>
<td>106.6</td>
<td>0.0</td>
<td>95.5</td>
<td>0.8</td>
</tr>
<tr>
<td>B2</td>
<td>24.5</td>
<td>154.6</td>
<td>9.05</td>
<td>463.2</td>
<td>12.2</td>
<td>24.2</td>
<td>0.0</td>
<td>3.4</td>
<td>101.5</td>
<td>0.0</td>
<td>95.5</td>
<td>0.8</td>
</tr>
<tr>
<td>B3</td>
<td>71.5</td>
<td>107.0</td>
<td>198.8</td>
<td>506.5</td>
<td>25.3</td>
<td>24.2</td>
<td>0.0</td>
<td>3.4</td>
<td>70.1</td>
<td>0.0</td>
<td>95.5</td>
<td>0.8</td>
</tr>
<tr>
<td>Group 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A6</td>
<td>-50.6</td>
<td>66.2</td>
<td>508.6</td>
<td>62.4</td>
<td>505.5</td>
<td>54.8</td>
<td>504.0</td>
<td>43.5</td>
<td>506.0</td>
<td>0.0</td>
<td>559</td>
<td>50.8</td>
</tr>
<tr>
<td>A7</td>
<td>-50.7</td>
<td>64.2</td>
<td>503.5</td>
<td>0.5</td>
<td>505.5</td>
<td>53.6</td>
<td>503.5</td>
<td>42.2</td>
<td>502.7</td>
<td>0.0</td>
<td>559</td>
<td>50.8</td>
</tr>
<tr>
<td>A8</td>
<td>-46.3</td>
<td>13.8</td>
<td>425.3</td>
<td>146.0</td>
<td>425.3</td>
<td>135.5</td>
<td>425.3</td>
<td>43.9</td>
<td>120.1</td>
<td>0.0</td>
<td>559</td>
<td>50.8</td>
</tr>
<tr>
<td>A9</td>
<td>-36.4</td>
<td>20.4</td>
<td>356.7</td>
<td>201.5</td>
<td>356.7</td>
<td>365.3</td>
<td>356.7</td>
<td>43.3</td>
<td>135.4</td>
<td>0.0</td>
<td>559</td>
<td>50.8</td>
</tr>
<tr>
<td>A10</td>
<td>-38.3</td>
<td>18.5</td>
<td>384.5</td>
<td>180.5</td>
<td>384.5</td>
<td>373.4</td>
<td>384.5</td>
<td>43.9</td>
<td>118.5</td>
<td>0.0</td>
<td>559</td>
<td>50.8</td>
</tr>
<tr>
<td>A11</td>
<td>-38.1</td>
<td>17.4</td>
<td>398.1</td>
<td>170.9</td>
<td>398.1</td>
<td>396.8</td>
<td>398.1</td>
<td>39.6</td>
<td>151.5</td>
<td>0.0</td>
<td>559</td>
<td>50.8</td>
</tr>
<tr>
<td>A12</td>
<td>-61.5</td>
<td>105.5</td>
<td>495.2</td>
<td>105.8</td>
<td>495.2</td>
<td>491.0</td>
<td>495.2</td>
<td>36.0</td>
<td>111.0</td>
<td>0.0</td>
<td>559</td>
<td>50.8</td>
</tr>
<tr>
<td>A13</td>
<td>-53.1</td>
<td>15.0</td>
<td>532.4</td>
<td>156.5</td>
<td>532.4</td>
<td>532.4</td>
<td>532.4</td>
<td>43.9</td>
<td>136.1</td>
<td>0.0</td>
<td>559</td>
<td>50.8</td>
</tr>
<tr>
<td>A14</td>
<td>-55.5</td>
<td>22.6</td>
<td>548.5</td>
<td>20.5</td>
<td>548.5</td>
<td>547.0</td>
<td>548.5</td>
<td>39.6</td>
<td>136.1</td>
<td>0.0</td>
<td>559</td>
<td>50.8</td>
</tr>
<tr>
<td>A15</td>
<td>-57.2</td>
<td>0.0</td>
<td>590.0</td>
<td>0.0</td>
<td>590.0</td>
<td>590.0</td>
<td>590.0</td>
<td>43.9</td>
<td>136.1</td>
<td>0.0</td>
<td>559</td>
<td>50.8</td>
</tr>
<tr>
<td>Group 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>A12</td>
<td>-51.1</td>
<td>23.1</td>
<td>510.5</td>
<td>160.5</td>
<td>510.5</td>
<td>510.5</td>
<td>510.5</td>
<td>43.9</td>
<td>136.1</td>
<td>0.0</td>
<td>559</td>
<td>50.8</td>
</tr>
<tr>
<td>A13</td>
<td>-62.3</td>
<td>41.4</td>
<td>584.5</td>
<td>131.5</td>
<td>584.5</td>
<td>584.5</td>
<td>584.5</td>
<td>54.8</td>
<td>151.5</td>
<td>0.0</td>
<td>559</td>
<td>50.8</td>
</tr>
<tr>
<td>A14</td>
<td>-46.2</td>
<td>20.6</td>
<td>485.5</td>
<td>185.5</td>
<td>485.5</td>
<td>485.5</td>
<td>485.5</td>
<td>43.9</td>
<td>136.1</td>
<td>0.0</td>
<td>559</td>
<td>50.8</td>
</tr>
<tr>
<td>A15</td>
<td>-48.4</td>
<td>27.8</td>
<td>485.5</td>
<td>185.5</td>
<td>485.5</td>
<td>485.5</td>
<td>485.5</td>
<td>43.9</td>
<td>136.1</td>
<td>0.0</td>
<td>559</td>
<td>50.8</td>
</tr>
<tr>
<td>A16</td>
<td>-56.9</td>
<td>20.5</td>
<td>522.1</td>
<td>118.9</td>
<td>522.1</td>
<td>522.1</td>
<td>522.1</td>
<td>43.9</td>
<td>136.1</td>
<td>0.0</td>
<td>559</td>
<td>50.8</td>
</tr>
<tr>
<td>A17</td>
<td>-57.2</td>
<td>19.9</td>
<td>565.5</td>
<td>100.5</td>
<td>565.5</td>
<td>565.5</td>
<td>565.5</td>
<td>43.9</td>
<td>136.1</td>
<td>0.0</td>
<td>559</td>
<td>50.8</td>
</tr>
<tr>
<td>A18</td>
<td>-61.5</td>
<td>109.9</td>
<td>427.2</td>
<td>127.6</td>
<td>427.2</td>
<td>427.2</td>
<td>427.2</td>
<td>43.9</td>
<td>136.1</td>
<td>0.0</td>
<td>559</td>
<td>50.8</td>
</tr>
<tr>
<td>A19</td>
<td>-68.5</td>
<td>7.6</td>
<td>440.4</td>
<td>30.7</td>
<td>440.4</td>
<td>440.4</td>
<td>440.4</td>
<td>43.9</td>
<td>136.1</td>
<td>0.0</td>
<td>559</td>
<td>50.8</td>
</tr>
<tr>
<td>A20</td>
<td>-76.2</td>
<td>0.0</td>
<td>671.1</td>
<td>0.0</td>
<td>671.1</td>
<td>671.1</td>
<td>671.1</td>
<td>43.9</td>
<td>136.1</td>
<td>0.0</td>
<td>559</td>
<td>50.8</td>
</tr>
</tbody>
</table>
32.0, 117, −342.0, 99.8, 186.0, 362.3, 495, and 883.5, respectively. However, eliminating this model (A5) from the candidate set of group 1 and recalculating the weights, one can find that AICc and SICc select the simplest model in the set (model A1) with \( w = 0.892 \) (for AICc) and 0.754 (for SICc), respectively, followed by model A2 as the second model with \( w = 0.108 \) and 0.245, respectively.

For the candidate set of group 2, AICc and SICc opt for model B10 as the best model with weight values equal to 1.0. In other words, the AICc and SICc results indicate that the remaining models have relatively no empirical support just like the other criteria. Moreover, SICc nominates model B15 with the chance of 82.9% (i.e., \( w = 0.892 \)) as the best approximating model among the candidate models in group 3, whereas AICc tends to select model B15 as the best model with \( w = 1 \). Taking SICc into consideration, there is a 94.7% confidence that the best model in group 3 is one of the models B14 and B15. Also, model B15 has an ER equal to 7 compared to model 14, which means that model 15 is seven times more likely to be the best model than model 14.

AICu and KICc show the same performance in the selection of the best model in groups 1, 2, and 3 with no distinctive ranking of the models. In this regard, the simplest model, i.e., A1, from the candidate set of group 1 remains the best as chosen by AICu and KICc with \( w \) values equal to 0.935 and 0.925, respectively, whereas models B10 and B15 have been chosen by both criteria as the best model from the candidate sets of groups 2 and 3 with \( w \) values equal to 1 and 0.999, respectively. It is worth noting that the two mentioned criteria cannot be calculated for model A5.

From the candidate models of group 1, HICc selects model B4 as the best one with \( w = 0.957 \), followed by model B3 as the second model with \( w = 0.043 \). Thus, model B4 is 22 times more likely to be the best model than model B3 (ER = 22). Similar to other criteria, HICc nominates models B10 and B15 as the best models of groups 2 and 3, respectively. Note that HICc is a small-sample bias correction version of HIC. We can verify when \( N \) is larger than \( K \), \( \Delta \) values of these two criteria become closer, so both of them have selected the same model as the best one in groups 2 and 3. However, \( \Delta \) values of the two criteria are completely different for group 1’s models. Therefore, the best selected model is different (models B4 and B5 for HICc and HIC, respectively).

AICu and KICc have been recommended for moderate-to-large sample size \( N \) and are known as large-sample based selection criteria (Rao & Wu 2001; Cavanaugh 2004). One can verify that for group 1’s models with \( N/K \) between 1 and 9, where \( K \) is a moderate-to-large fraction of sample size, AICu and KICc have given the highest rank to the simplest model (model A1) that shows the dominance of the model complexity term. Consequently, these criteria do not perform accurately when \( N/K \) is low.

AICc and SICc, which respectively are small-sample bias correction versions of AIC and SIC, have been recommended where \( N/K \) is low (say \( N/K \) values less than 40 for AICc). For group 1’s models with \( N/K \) between 1 and 9, these two criteria have selected model A5 with the highest \( N/K \) (=0.9). As mentioned before, the number of parameters for model A5 is near \( N \), causing a large negative value for the penalty term. Eliminating this model (A5) from the candidate set of group 1 and recalculating the weights, one can find that AICc and SICc select the simplest model in the set (model A1) with the highest \( N/K \) (=9.0) showing the dominance of the model complexity term. Model A1 has only one parameter with the largest RMSE (7.46 m) and cannot be a good candidate for the best model. Other criteria have selected the model whose \( N/K \) is equal to 1.2 (model B5) and the dominant role of the factor of model fit. This model is the most complex model among the set of group 1’s models with RMSE equal to 0.61 m. Note that HICc has opted for model B4 whose RMSE is very close to the minimum error. The results show that CAIC, HICc, and RICc have unexpectedly opted for the most complex models of group 1 (model B4 and B5).

For group 2’s models with \( N/K \) varying from 11 to 108, all 12 criteria have chosen the model with the lowest values of both \( N/K \) and RMSE (model B10). In other words, for \( N = 216 \) and \( K \) ranging from 1 to 19, these criteria have selected the model with the highest complexity (model B10). According to the results illustrated in Table 3, one can infer that while \( N/K \) has varied from 3 to 72 for group 3’s models, all the criteria have selected the model with the lowest \( N/K \) (model B15). However, defining a new group containing models A6–A15 and B6–B15, one can verify that all the criteria tend to opt for the model
Figure 5 | Comparison of simulated and observed water levels at the observation wells 17, 7, and 5 between models B10 and B15 during the validation period. The target interval is ±1.0 m.
with the highest weight and the lowest $N/K$ (model B15), similar to the ranking of group 3’s models. It shows that when $N/K$ has varied from a moderate value to a large value, all 12 criteria have a tendency to overfit.

From an overall view, the results indicate that all the criteria evaluated in the present study have recommended models B15 and B10 as the best models in the one-step and the two-step calibration approaches, respectively. Model B10 has the lowest RMSE value from the candidate set of group 2’s models, and model B15 has the lowest RMSE value among all the candidate models. For both of them, the parameterization of HK and $S_\text{f}$ has been done by RPP and ZMs, respectively.

We finally tested the selected models based on a monthly data set of observed water levels from October 2001 to September 2006. A meaningful difference is observed in their performances as the RMSE values are equal to 0.68 m and 0.97 m for models B10 and B15, respectively. Figure 5 displays the simulated and observed water levels at three observation wells during the validation period. The location of these observation wells is shown in Figure 1. One can see that monthly water level fluctuations have been simulated well just in model B10, and model B15 has failed to simulate the fluctuations accurately.

**CONCLUSIONS**

This study elaborated on how to rank different conceptualizations of a groundwater system by using appropriate selection techniques and criteria that make a balance between the complexity and accuracy features of models.

We used a multi-modeling approach with regard to models’ parameterization and calibration. Different candidate models were generated depending on the parameterization method of HK by either ZM (type A) or RPP method (type B), the number of zones in the ZM or number of pilot points in the RPP method and the calibration approach (one-step or two-step). This resulted in 30 different models (Table 2) that were then evaluated and compared against 12 discriminating selection criteria. Some of the criteria, i.e., AIC, AICc, SIC, and HIC had been widely used in groundwater modeling, and the others, i.e., AICu, CAIC, SICc, DIC, HICc, RICc, AIC3, and KICc had not. The results obtained from the use of several models and selection techniques in modeling of the Dehloran groundwater system in Iran, in which a SWAT-based distributed hydrological model estimates the spatio-temporal patterns of recharge, indicate that:

1. In the two-step calibration approach, the models with high ranks in the transient mode are not necessarily the best corresponding models in the steady-state calibration approach. This situation was observed for AICc, AICu, SICc, HICc, and KICc. This reveals that low rank models in the first step of calibration should not be disregarded from further consideration, and the selection procedure must be done at the end of the second calibration step.

2. Models B10, calibrated by the two-step approach using RPP and ZMs of HK and $S_\text{f}$ parameterization, respectively, and B15, calibrated by the one-step approach using the same parameterization methods, have been better than other candidate models based on calibration results. However, the validation results reveal that the best model calibrated by the two-step approach has been far superior than the one calibrated by the one-step approach.

3. Regarding the number of observations, AICc and SICc criteria could be very sensitive to the number of calibration parameters for some of the models. This is the case for HICc, AICc, KICc, and RICc too because the penalty terms of these criteria are not continuous functions of $K$ (number of parameters), so they may change abruptly at some certain points, approximately around $N = K$ point. Therefore, more care should be paid when using these criteria.

4. The ratio of the number of total observations ($N$) to the number of parameters ($K$) strongly affects the performance of the selection criteria. AICc had been recommended when $N/K$ was less than 40 (Burnham & Anderson 2002; Poeter & Anderson 2005; Hill & Tiedeman 2007). In this study, AICc has been unable to determine the best model when $N/K$ varied from 1 to 9. In addition, it had been reported that AICu had provided better model choices than AICc for moderate-to-large sample sizes $N$ without defining any range (Rao & Wu 2001). Our results have indicated that AICu may not work well enough for the values of $N/K$ between 1 and 9.
ACKNOWLEDGEMENT

This paper has been partially supported by the research grant No. 40/758 provided by Research Grants Program of Amirkabir University of Technology to the corresponding author.

REFERENCES


Hill, M. C. 2006 The practical use of simplicity in developing groundwater models. Ground Water 44 (6), 775–781.


First received 14 August 2014; accepted in revised form 10 April 2015. Available online 4 June 2015.