The characteristics of probability distribution of groundwater model output based on sensitivity analysis

Xiankui Zeng, Jichun Wu, Dong Wang and Xiaobin Zhu

ABSTRACT

The probability distribution of groundwater model output is the direct product of modeling uncertainty. In this work, we aim to analyze the probability distribution of groundwater model outputs (groundwater level series and budget terms) based on sensitivity analysis. In addition, two sources of uncertainties are considered in this study: (1) the probability distribution of model's input parameters; (2) the spatial position of observation point. Based on a synthetical groundwater model, the probability distributions of model outputs are identified by frequency analysis. The sensitivity of output's distribution is analyzed by stepwise regression analysis, mutual entropy analysis, and classification tree analysis methods. Moreover, the key uncertainty variables influencing the mean, variance, and the category of probability distributions of groundwater outputs are identified and compared. Results show that mutual entropy analysis is more general for identifying multiple influencing factors which have a similar correlation structure with output variable than a stepwise regression method. Classification tree analysis is an effective method for analyzing the key driving factors in a classification output system.

Key words | classification tree analysis, frequency analysis, groundwater modeling, mutual entropy analysis, probability distribution, sensitivity analysis

Xiankui Zeng
Jichun Wu (corresponding author)
Dong Wang
Xiaobin Zhu
Key Laboratory of Surficial Geochemistry, Ministry of Education,
Department of Hydrosciences,
School of Earth Sciences and Engineering,
State Key Laboratory of Pollution Control and Resource Reuse,
Nanjing University,
Nanjing, 210093,
China
E-mail: *jcwu@nju.edu.cn*

INTRODUCTION

Groundwater modeling and prediction are influenced by many factors from the surface to underground. The uncertainty of groundwater model outputs stems from a number of factors including incomplete model structure, incorrect boundary conditions, and aquifer parameters (Hassan *et al.* 2008; Wu *et al.* 2011; Zhang *et al.* 2012; Gungor & Goncu 2013; Zeng *et al.* 2013). Data scarcity and observation errors enhance the difficulty in handling simulation uncertainty (Hassan *et al.* 2008; Mpimpas *et al.* 2008; Wang *et al.* 2009). In recent years, a number of studies have been developed to assess the uncertainties on groundwater model outputs. Moreover, these studies focus on the uncertainty assessments by referring uncertainty sources such as hydrogeological parameters, conceptual model, and scenario (Blasone *et al.* 2008; Hassan *et al.* 2008; Ye *et al.* 2010; Hashemi *et al.* 2013; Morway *et al.* 2013).

Uncertainty analysis of groundwater models is often implemented in a probability statistical framework (Blasone *et al.* 2008; Hassan *et al.* 2008). The results are generally expressed as the probability distributions of outputs of interest (e.g., groundwater level, boundary flux, solute concentration). The uncertainty of a random variable can be described by its characteristics of probability distribution, which include probability density function (PDF) and numerical characteristics (e.g., mean and variance). The location, range, and shape of a random variable's distribution are determined by the probability distribution. The sensitivity analysis of groundwater output's probability distribution is primarily aimed at identifying two types of influencing factors. One is the factor affecting the numerical characteristics of a random variable, the other is the driving factor which leads the output variable to obey a specific PDF.

doi: 10.2166/hydro.2013.106

For the uncertainty analysis of groundwater simulation, in general, we are interested in the probability distribution of model output. However, little attention has been devoted to the influencing factors of model output's distribution in previous studies. In this paper, we focus on the sensitivity of groundwater model output's probability distribution for two sources: (1) the probability distribution of the input parameters; (2) the spatial position of observation point.

Frequency analysis is a technique which has been extensively used in hydrologic uncertainty issues (Lang *et al*. 2010; Neppel *et al*. 2010), such as the design of flood control and risk management. The observation series is used to fit an alternative PDF, and then the variable's distribution uncertainty is analyzed statistically (Smakhtin 2001; Katz *et al.* 2002). Generally, there are three basic procedures for frequency analysis: (1) selecting a suitable PDF for data series; (2) parameter estimation for the selected PDF; (3) uncertainty assessment for the data series (Onoz & Bayazit 1995). Herein, how to select a suitable PDF is the key problem for a frequency analysis. According to Mcmahon & Srikanthan (1981), Haktanir (1992), Onoz & Bayazit (1995), and Vogel *et al.* (1993), there is not a universal applicable rule to select the best PDF, and the qualified PDF should be selected based on effective comparison and testing.

For the complicated groundwater model, it is hard to describe the influences of model inputs on outputs directly by mathematic model. Sensitivity analysis provides an effective framework for unraveling the relationship between the input variables and outcomes. In general, the studying object is the direct model output, such as hydraulic head (Rojas *et al*. 2009; Mazzilli *et al*. 2010) and solute concentration (Huysmans *et al*. 2006; Zhang *et al*. 2009). The influencing factors of output variable can be identified by sensitivity analysis. In this study, the research object is not the direct model output, but the probability distribution of output. The importance of this kind of influencing factor can be regarded as another form of sensitivity to model output. Furthermore, recognizing the distribution characteristics of model output will help in identifying groundwater modeling uncertainty, improving model structure, and providing feedback for data collecting activities relating to model uncertainty analysis.

A synthetic groundwater model was built for producing groundwater outputs. The outputs of the model include groundwater levels series (GLS) and groundwater budget terms. The suitable PDFs of outputs were selected by the Kolmogorov–Smirnov test. After that, the stepwise regression and mutual entropy analysis were used to identify the influencing factors of the first two moments of GLS (mean and variance). In addition, mutual entropy analysis is a reliable sensitivity analysis method based on information theory, which is compared with stepwise regression analysis. Finally, for the sensitivity analysis of classification output system, classification tree analysis was used to identify the driving factors that lead the GLS to obey a specified distribution.

The main results of this study were obtained from a synthetic groundwater model. This groundwater model is simple compared to a real groundwater system. Therefore, the research results can be regarded as a mathematical exploration into the characteristics of probability distribution of groundwater model outputs. Some conclusions need further confirmation in the real field. Nevertheless, the use of a real groundwater model is not easy for such analysis, because observations are often limited in the number and length of a data series.

In the following sections, the methods used for this research are described. Then, a synthesized groundwater flow model is presented. In the results and discussion section, we describe the characteristics of probability distribution of model output. Finally, the main conclusions drawn from the analysis are provided.

METHODS

Parameter estimation and goodness of fit test

Seven functions were chosen as the alternative probability distribution functions to fit the outputs of groundwater model. They were normal, log-normal, 2-parameter gamma, log-2-parameter gamma, Pearson type III, log-Pearson type III, and uniform distribution, respectively. The methods used for parameter estimation have been illustrated in many papers and will not be provided here. Readers can obtain detailed derivation processes by referring to Chen *et al*. (2002), Ross (2004), Singh & Singh (1985a, b), and Sun & Zheng (2006).

The Kolmogorov–Smirnov test (Melo *et al.* 2009; Wang & Wang 2010) is a convenient method of a hypothesis test by comparing the statistic value with the critical value at a specified confidence level. The statistic value is evaluated by comparing the proposed PDF with the empirical distribution function constructed based on samples. The Kolmogorov–Smirnov test is a standard procedure of goodness of fit test, and it will not be described here.

Stepwise regression analysis

Stepwise regression analysis is a common approach for global sensitivity analysis. The basic idea for regression analysis is to fit the input and output variable with a linear regression model (Pappenberger *et al.* 2008; Mishra *et al.* 2009). The model generated at every step is tested to ensure that all the regression variables are important to the model. The *t*-test measuring the difference between samples and the regression model is applied to test the importance of a variable. In addition, if some variables are found to be insignificant, then the most insignificant variable is removed from the model. Moreover, the stepwise regression process will continue until each variable in the regression model becomes significant and the variables outside of the model are insignificant (Mishra *et al.* 2009; Bergante *et al.* 2010; Zeng *et al.* 2012). After that, the uncertainty importance of input variable can be defined as standardized regression coefficient (SRC):

$$\mathrm{SRC} = \frac{b_j \sigma(x_j)}{\sigma(y)} \tag{1}$$

where y is the output variable, x_j is the input variable numbered by j, $\sigma(x_j)$, $\sigma(y)$ are the standard deviations of x_j and y, respectively, b_j is the regression coefficient of x_j.

Mutual entropy analysis

The distribution character of data set (X, Y) can be described using contingency tables. For the contingency tables' rows, the label denotes the input variable x, and the range is divided into i equal-width intervals. For the contingency tables' columns, the label denotes the output variable y, and the range is divided into j equal-width intervals. The number in each contingency table is a nonnegative integer which represents the number of observed events satisfying the joint conditions of row and column.

The probability of the state with input variable x_i and output variable y_i is $p_{ij} = N_{ij}/N$, where N_{ij} is the value of the contingency table at i-th row and j-th column, and N is the number of samples. In addition, $N_{i.}$ is the cumulative number of samples in the i-th interval of x for the whole range of y, and $N_{.j}$ denotes the cumulative number of samples in the j-th interval of y for the whole range of x. Consequently, when considering the state x_i only, the probability can be written as $p_{i.} = N_{i.}/N$, and the probability of outcomes only with the state y_j is given by $p_{.j} = N_{.j}/N$ (Mishra *et al.* 2009).

The entropy of a variable represents the amount of average information. According to information theory, the entropies of variable x, y, and (x, y) are defined as follows:

$$H(x) = -\sum_i p_{i.} \ln p_{i.}; \; H(y) = -\sum_j p_{.j} \ln p_{.j} \tag{2}$$

$$H(x, y) = -\sum_i \sum_j p_{ij} \ln p_{ij} \tag{3}$$

In information theory, the mutual information of two variables is a quantity that measures the mutual dependence of two variables. The mutual entropy between x and y is described as the reduction in the uncertainty of y due to the information of x, which can be given by:

$$I(x, y) = H(x) + H(y) - H(x, y) = \sum_i \sum_j p_{ij} \ln \frac{p_{ij}}{p_{i.} p_{.j}} \tag{4}$$

In mutual entropy method, the uncertainty importance of input variables on output variable is indicated by two indicators: uncertainty coefficient (U) and R statistic (R) (Mishra *et al.* 2009; Zeng *et al.* 2012):

$$U(x, y) = 2\left[\frac{I(x, y)}{H(x) + H(y)}\right] \tag{5}$$

$$R(x, y) = [1 - \exp\{-2I(x, y)\}]^{1/2} \tag{6}$$

These two measures take values in the range [0, 1], U (or R) is 0 if x and y are independent, and it takes 1 if x is completely related to y.

Classification tree analysis

Sensitivity analysis techniques such as stepwise regression, regionalized sensitivity analysis (Pappenberger *et al.* 2008), and mutual entropy analysis are useful for identifying important influencing factors if the study object is a continuous variable. When the problem relates to binary outcomes such as 'right' vs. 'wrong', 'yes' vs. 'no', the classification tree method provides a more efficient framework for identifying the factors driving the result into particular categories (Mishra *et al.* 2003; Englehart & Douglas 2010; Esther *et al.* 2010; MacQuarrie *et al.* 2010).

The fundamental target for constructing a classification tree model is searching for a classifying rule. The output is classified by a series of splits based on splitting variables. Each split is determined by the appropriate classifier. Thus, the following two steps are essential for constructing a classification tree: (1) selecting an appropriate splitting variable and determining the split point; (2) deciding when to continue splitting or to declare splitting termination.

The split can be defined by several principles, such as maximum information gain (InfoGain), maximum impurity reduction, and maximum reduction in deviance (Mishra *et al.* 2003; Myles *et al.* 2004). The InfoGain index based on information entropy theory was applied to construct a classification tree in this study. The outputs are classified into subspaces by selecting a splitting point of the splitting variable. This implies the complicated and disordered outputs are arranged and sorted with higher order within subspaces. Therefore, the uncertainty of output variable is reduced by acquiring information.

A classification tree is built on two types of nodes: branch nodes and leaf nodes. Each branch node is the parent of two children branch nodes, and the leaf node is the endpoint of the tree.

The uncertainty or information entropy of output variable y in node t is defined as:

$$Info = -\sum_{j}\left(\frac{N_j(t)}{N(t)}\right)\ln\left(\frac{N_j(t)}{N(t)}\right) \tag{7}$$

where $N_j(t)$ denotes the number of samples belonging to the class j at node t, and $N(t)$ is the number of samples at node t.

Assuming the splitting variable X, with n samples ordered by magnitude, the amount of alternative split points of X is n-1 by choosing the midpoint of two adjacent samples. The point that maximizes the information gain or minimizes the uncertainty of outputs is selected. InfoGain is calculated by the equation (Myles *et al.* 2004):

$$InfoGain = Info(parent) - \sum_{k}(p_k)Info(child_k) \tag{8}$$

where *parent* denotes the space before splitting, $child_k$ denotes the subspace after splitting, and p_k is the ratio of the samples which passed into the k-th subspace. The purity of a space describing the distribution of samples' types is expressed as follows:

$$purity = \sum_{j} p_j^2;\ p_j = \frac{N_j(t)}{N(t)} \tag{9}$$

where p_j is the proportion of samples belonging to class j.

The classification tree is constructed by the successive selection of splitting points. It is beneficial to set up some constraint for preventing excessive splitting. If the number of samples in a subspace below the minimum value, or the purity of samples in a subspace is higher than the maximum value specified by the user, the splitting is terminated at that node. Furthermore, a classification tree can be optimized by pruning and reconstruction, which acquires a balance between the complexity and classification precision. After the classification tree is constructed, the sensitivities of splitting variables can be simply determined by comparing the order used to classify outputs (Mishra *et al.* 2003).

IMPLEMENTATION OF METHODS

Description of the synthesized model

For the purpose of frequency analysis, we constructed a synthetic three-dimensional steady-state groundwater model (Rojas *et al.* 2009) (Figure 1). The model domain is 5,000 m in the x direction, 3,000 m in the y direction, and 53 m in the z direction (thickness). The model area is a rectangle (5,000 m by 3,000 m) and discretized into 25 m by

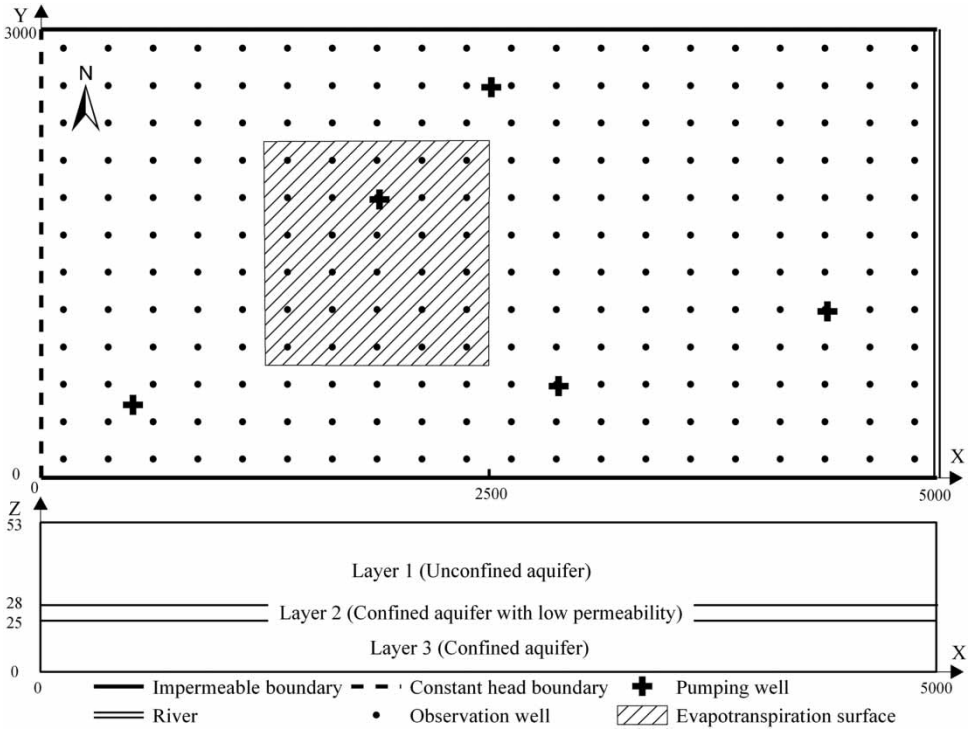

Figure 1 | A schematic diagram of synthetic groundwater model.

25 m grid cells. Five pumping wells and 240 observation points are placed at the confined aquifer. The upper layer is 25 m in thickness and is modeled as an unconfined aquifer. The lower layer is 25 m in thickness and is confined. The upper and lower layer is separated by a 3-m confining bed by neglecting its storage capacity. The three model layers were assumed to be horizontal in extension. The hydraulic conductivity distribution within each aquifer is heterogeneous, and the hydraulic conductivity field within each layer is assumed to be statistically stationary.

Model parameters

Model layers are assumed to be homogeneous statistically with a constant mean of hydraulic conductivity K. Smaller-scale variability is represented using the theory of random space functions. In addition, an isotropic exponential covariance function is used to describe the K fields of layers. The spatial distribution of hydraulic conductivity is generated using the direct Fourier transform method (Robin *et al.* 1993). The spatial structure parameters of ln$K$ for different layers are presented in Table 1.

Table 1 | Spatial correlation parameters of hydraulic conductivity K for the layers of synthetic groundwater model

	Parameter		
Layer	Mean of K (m/d)	Variance of ln$K$	Correlation length of ln$K$ (m)
1	5.0	2.0	80
2	0.1	0.5	80
3	5.0	2.0	80

Boundary conditions set up

As shown in Figure 1, for the model aquifers, two impermeable boundary conditions are specified along the south and north boundaries. Along the west boundary, a constant head boundary condition is imposed. The east side of the domain is bounded by a 20 m-wide river, and the river level is 40 m. The riverbed's thickness is 2 m, and the elevation at the bottom of the riverbed is 35 m. Furthermore, sources and sinks in the model include recharge from precipitation, discharge from pumping and evapotranspiration. The top surface of unconfined aquifer receives the precipitation recharge uniformly, and the model bottom is an

impermeable boundary. In addition, the pumping wells and observation wells are only screened at layer 3. An evapotranspiration zone, delineated by a rectangle in the left side of the study area, is defined with an evapotranspiration surface elevation at 51 m, and the extinction depth is set as 5 m.

Then, the unknown model parameters including the water level of constant head boundary, the conductance of river bed, precipitation rate, maximum evapotranspiration rate, and pumping rate are defined in specified ranges (Table 2). In addition, the conductance of riverbed represents the interconnection between river and unconfined aquifer, which is calculated as follows (Harbaugh 2005):

$$CRiv = \frac{KRiv * l * w}{m} \tag{10}$$

where $CRiv$ is the conductance of riverbed, $KRiv$ is the vertical hydraulic conductivity of riverbed, l is the length of reach, w is the width of river, and m the thickness of riverbed.

The probability distribution of groundwater model output is influenced by input parameters. Therefore, two conditions that the input parameters follow, uniform and normal distributions, are both considered in this study. In addition, the range of uniform distribution is consistent with interval of corresponding normal distribution. The parameters of these two distributions are shown in Table 2.

Monte Carlo simulation

The numerical model of synthesized groundwater flow system is built using MODFLOW-2005 (Harbaugh 2005).

Table 2 | Probability distributions of model parameters

	Uniform distribution		Normal distribution	
Model parameter	**Minimum**	**Maximum**	**Mean**	**Variance**
Precipitation rate (m/d)	6.0×10^5	6.0×10^4	3.3×10^4	7.1×10^5
Evapotranspiration rate (m/d)	5.0×10^4	5.0×10^3	2.75×10^3	5.92×10^4
Constant head (m)	47.0	52.0	49.5	0.6579
Conductance of riverbed (m^2/d)	10.0	500.0	255.0	64.4737
Pumping rate (m^3/d)	500.0	3000.0	1750.0	328.9474

The Monte Carlo simulation procedure involves two parts (part I and part II).

Part I

Part I includes the following steps:

1. Generating model mesh, setting the initial head condition, the positions of pumping wells and observation points, etc.
2. Setting the hydraulic conductivity K of model layers. Based on the mean and the covariance function of ln$K$ (Table 1), the random fields of K are generated by the direct Fourier transform method.
3. Setting the boundary conditions, including precipitation rate, maximum evapotranspiration rate, water head of constant head boundary, conductance of riverbed, and pumping rate. A boundary condition is assigned a value by sampling uniformly from the corresponding range (Table 2).
4. Running the established model and collecting the outputs of groundwater model. The outputs include the groundwater levels of observation points in layer 3, the inflow from constant head boundary and precipitation, the outflow from well pumping, evapotranspiration process, and river boundary.
5. Repeating step 2 to step 4 500 times.
6. Conducting frequency analysis for groundwater model outputs. The data series used for frequency analysis is constructed by the output of every realization, e.g., the groundwater levels of an observation point from 1st to 500th realization. Therefore, each data series has 500 samples. The data series include 240 GLS and five groundwater budget series. The procedure of frequency analysis can be summarized as two steps: (1) parameter estimation for each alternative PDF; (2) taking the Kolmogorov–Smirnov test for each PDF. If all the alternative PDFs have poor performance (cannot pass through the Kolmogorov–Smirnov test, and the significance level α was set to 0.05 in this study), we will mark the GLS as an unknown PDF.

Part II

The procedure of part II is the same as that of part I, except for step 3. In this part, step 3, a boundary condition is assigned a value by sampling from corresponding normal distribution.

RESULTS AND DISCUSSION

Figure 2 shows the frequency distribution of parameters' samples which are sampled from uniform and normal distributions, respectively.

Frequency analysis

The outputs of groundwater model are tested for each alternative PDF by Kolmogorov–Smirnov test, and the significance level is 0.05. The numbers of GLS which obey normal, log-normal (Log-nor), 2-parameter gamma (G2), log-2-parameter gamma (Log-G2), Pearson type III (P3), log-Pearson type III (Log-P3), uniform, and unknown distribution are denoted as n_i ($i = 1,2,\ldots,8$) in order. After that, the ratio for each PDF was calculated as:

$$ratio_i = n_i/240 \tag{11}$$

As shown in Figure 3, the PDF of GLS is strongly influenced by the probability distribution of model input parameters. When the input parameters are sampled from uniform distribution, although a majority of GLS obey unknown distribution (Figure 3(a)), the rest of GLS obey uniform distribution nearly. Moreover, when the input parameters are sampled from normal distribution, it is obvious that most of the GLS obey normal distribution (Figure 3(b)).

The groundwater budget terms include the inflows from constant head boundary (InCH) and precipitation (InPre), and outflows from river leakage (OutRiv), evapotranspiration (OutEva), and pumping (OutPum). Obviously, the probability distributions of InPre and OutPum are fully controlled by model input parameters (precipitation rate and pumping rate). Figure 4 shows the frequency distributions of InCH, OutRiv, and OutEva. When input parameters are sampled from uniform distribution, none of the budget terms can pass the Kolmogorov–Smirnov test (Figures 4(a)–4(c)). Moreover, the distributions of these budget terms are significantly different from uniform distribution. By contrast, all the budget terms have passed the Kolmogorov–Smirnov test as normal distribution when input parameters are sampled from normal distribution (Figures 4(d)–4(f)).

Stepwise regression analysis

Figure 3 shows that only a part of GLS obeys a specified PDF. The observed GLS show different characteristics of

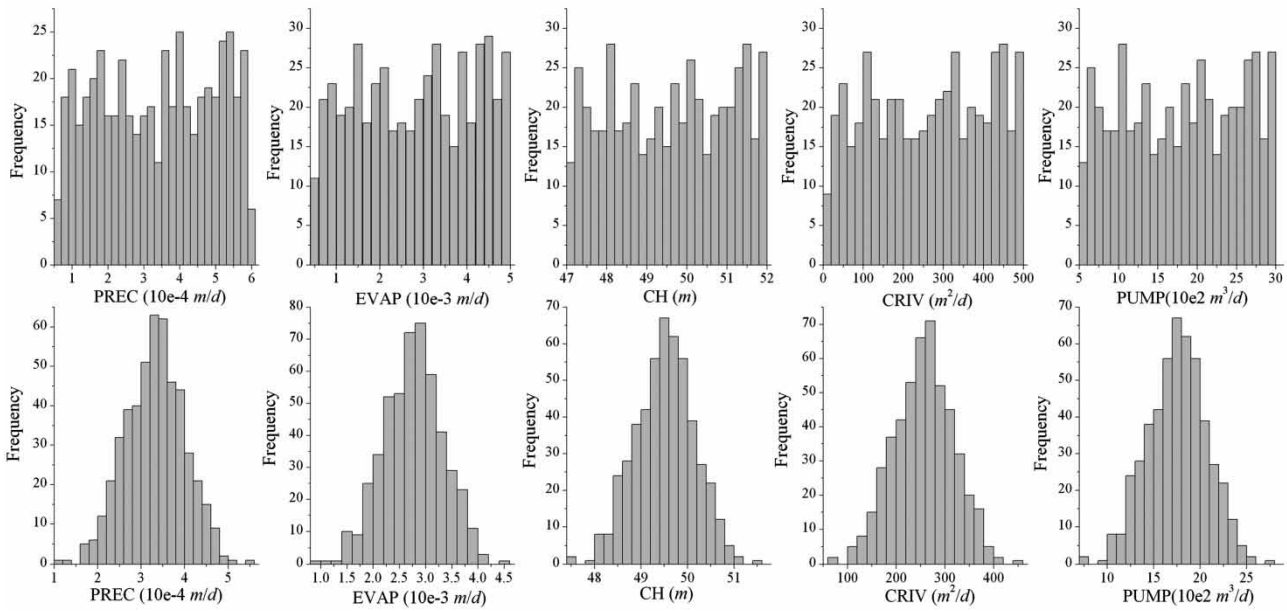

Figure 2 | The frequency distributions of precipitation rate (PREC), evapotranspiration rate (EVAP), constant head (CH), conductance of riverbed (CRIV), and pumping rate (PUMP). First and second rows represent these parameters are sampled from uniform and normal distributions, respectively.

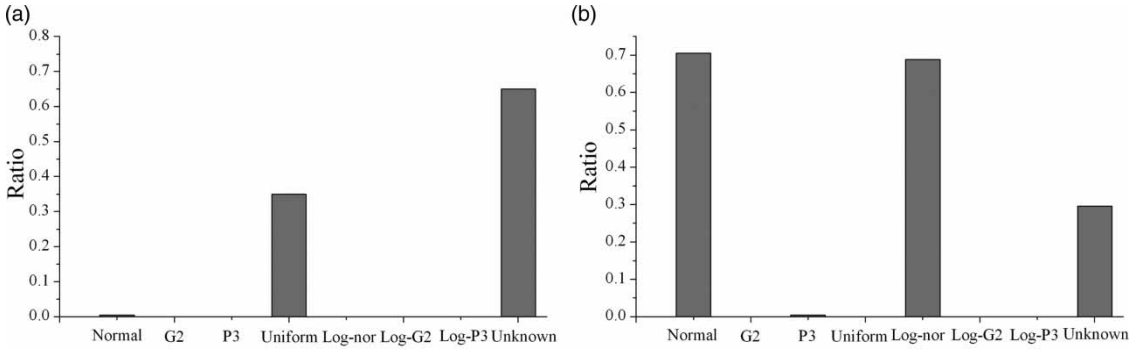

Figure 3 | The ratios of GLS which obey normal, G2, P3, uniform, Log-nor, Log-G2, Log-P3, and unknown distribution. (a) and (b) denote input parameters are sampled from uniform and normal distributions, respectively.

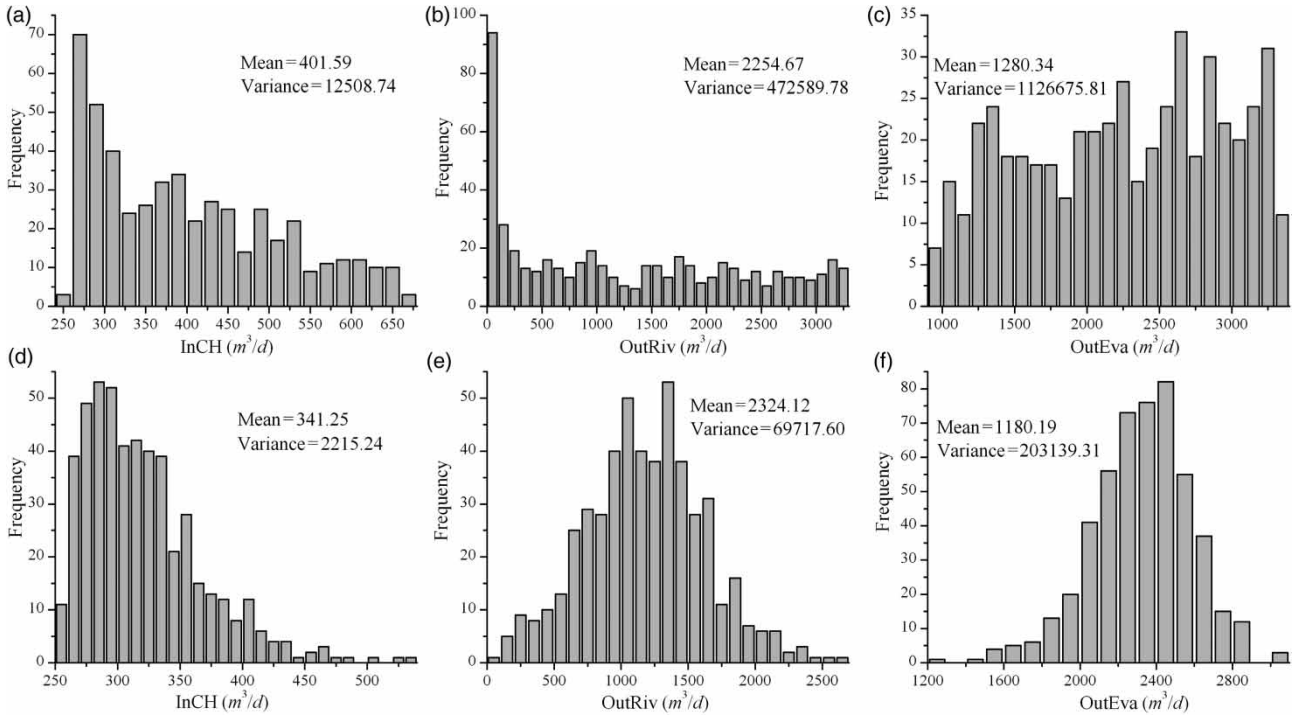

Figure 4 | Frequency distributions of inflow from constant head boundary (InCH), outflows from river leakage (OutRiv) and evapotranspiration (OutEva). The plots (a), (b), (c) and the plots (d), (e), (f) denote input parameters are sampled from uniform and normal distributions, respectively.

probability distributions among observation points. The probability distribution of GLS is influenced by the spatial position of an observation point. Thus, the stepwise regression analysis is used to identify the key factors of the mean and variance of GLS. The input variables of regression model are the distances of an observation point from surrounding model boundaries. They are the distances of an observation point from northern boundary, river boundary, southern boundary, constant head boundary, the nearest pumping well, evapotranspiration area, and the average distance from five pumping wells. The variables are listed in Table 3 and numbered from 1 to 7, all of them are normalized before regression analysis.

As shown in Figure 5, as the input parameters are sampled from uniform and normal distribution, respectively, the sensitivities of influencing factors are almost identical for the mean of GLS, e.g., Figure 5(a) vs. Figure 5(b). However, the influences of regression variables on the variance

Table 3 | Input variables of stepwise regression model and their numbers

Variable	No.
Distance from an observation point to northern boundary (D1)	1
Distance from an observation point to river boundary (D2)	2
Distance from an observation point to southern boundary (D3)	3
Distance from an observation point to constant head boundary (D4)	4
Distance from an observation point to the nearest pumping well (D5)	5
Average distance from an observation point to five pumping wells (D6)	6
Distance from an observation point to evapotranspiration area (D7)	7

of GLS are slightly different for these two distributions, such as Figure 5(c) vs. Figure 5(d).

For the regression analysis of the mean of GLS, four variables (D2, D5, D6, and D7) passed into the regression model. The variable with the largest sensitivity is D2 (the regression coefficient is about 0.97). For the regression analysis of the variance of GLS, four variables (D2, D5, D6, and D7) also passed into the regression model. The variable with the largest sensitivity is also D2 (the regression coefficient is about 0.80). Therefore, the mean and variance of GLS are affected similarly by the regression variables. They are both significantly influenced by the distance from river boundary (D2), and other regression variables have very low influences relative to D2. Furthermore, the regression coefficient of D2 in the mean model is larger than that in the variance model, e.g., Figure 5(a) vs. Figure 5(c), Figure 5(b) vs. Figure 5(d). Thus, the mean of GLS is more dependent on the distance from river boundary than the variance of GLS. In addition, the average distance from five pumping wells (D6) is inversely related with the mean and variance of GLS.

Mutual entropy analysis

Stepwise regression analysis is restricted in monotonic linear issues, and mutual entropy analysis is capable of treating the complicated non-monotonic relationship between output and input variables. The same as for stepwise regression analysis, the input variables are also listed in Table 3, and the output variables are the mean and variance of GLS. Tables 4–7 display the contingency tables of mutual entropy analysis.

Figure 6 shows the results of mutual entropy analysis. Similar to the results of stepwise regression analysis, the sensitivities of input variables are similar for the mean and variance of GLS. The most important influencing factors for the mean and variance of GLS are the distances from an observation point to river and constant head boundaries (D2 and D4). In addition, for the mean of GLS, the variables with the weakest sensitivity are the distances from an observation point to northern and southern boundaries (D1 and D3). For the variance of GLS, the distance from an observation point to evapotranspiration area (D7) holds the smallest sensitivity. Nevertheless, the index values of D1, D3, and D7 are very close for the mean and variance of

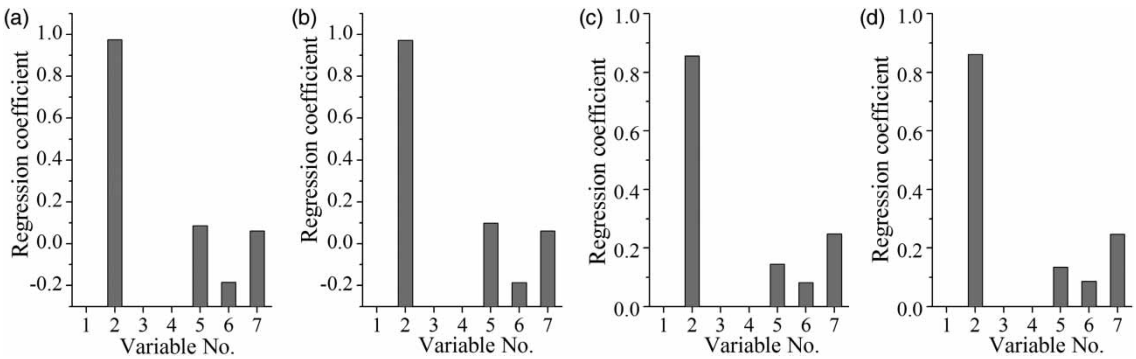

Figure 5 | The regression coefficients of the entered variables in stepwise regression analysis. The plots (a), (b) and the plots (c), (d) denote output variables are the means and variances of GLS, respectively. The plots (a), (c) and the plots (b), (d) indicate input parameters are sampled from uniform and normal distributions, respectively.

Table 4 | Contingency tables when *Y* (labeled by column) is the mean of GLS, and model parameters are sampled from uniform distribution

D1				D2				D3				D4				D5				D6				D7			
9	9	21	21	36	24	0	0	9	11	17	23	0	0	0	60	9	5	36	13	0	8	66	11	12	14	55	43
9	9	24	18	0	16	44	0	9	11	21	19	0	0	39	21	14	20	44	36	4	24	17	36	0	0	0	0
9	11	21	19	0	0	39	21	9	9	24	18	0	16	44	0	8	13	4	23	20	8	0	27	12	14	19	25
9	11	17	23	0	0	0	60	9	9	21	21	36	24	0	0	5	2	0	9	12	0	0	7	12	12	9	13

Table 5 | Contingency tables when *Y* (labeled by column) is the variance of GLS, and model parameters are sampled from uniform distribution

D1				D2				D3				D4				D5				D6				D7			
9	31	8	12	40	20	0	0	11	27	14	8	0	4	18	38	24	27	4	8	20	65	0	0	34	61	17	12
21	26	4	9	1	57	2	0	19	28	4	9	19	31	10	0	23	64	18	9	8	39	28	6	0	0	0	0
19	28	4	9	19	31	10	0	21	26	4	9	1	57	2	0	8	20	8	12	20	8	2	25	13	32	11	14
11	27	14	8	0	4	18	38	9	31	8	12	40	20	0	0	5	2	0	9	12	0	0	7	13	19	2	12

Table 6 | Contingency tables when *Y* (labeled by column) is the mean of GLS, and model parameters are sampled from normal distribution

D1				D2				D3				D4				D5				D6				D7			
9	9	19	23	36	24	0	0	9	9	17	25	0	0	0	60	9	4	36	14	0	5	66	14	12	12	54	46
9	9	23	19	0	12	48	0	9	9	22	20	0	0	33	27	14	17	42	41	4	23	15	39	0	0	0	0
9	9	22	20	0	0	33	27	9	9	23	19	0	12	48	0	8	13	4	23	20	8	0	27	12	12	19	27
9	9	17	25	0	0	0	60	9	9	19	23	36	24	0	0	5	2	0	9	12	0	0	7	12	12	8	14

Table 7 | Contingency tables when *Y* (labeled by column) is the variance of GLS, and model parameters are sampled from normal distribution

D1				D2				D3				D4				D5				D6				D7			
14	29	6	11	41	19	0	0	11	31	11	7	0	6	20	34	30	21	5	7	37	48	0	0	51	47	16	10
25	23	3	9	5	55	0	0	28	20	5	7	32	23	5	0	34	60	13	7	9	47	22	3	0	0	0	0
28	20	5	7	32	23	5	0	25	23	3	9	5	55	0	0	9	21	7	11	20	8	3	24	13	36	9	12
11	31	11	7	0	6	20	34	14	29	6	11	41	19	0	0	5	2	0	9	12	0	0	7	14	20	0	12

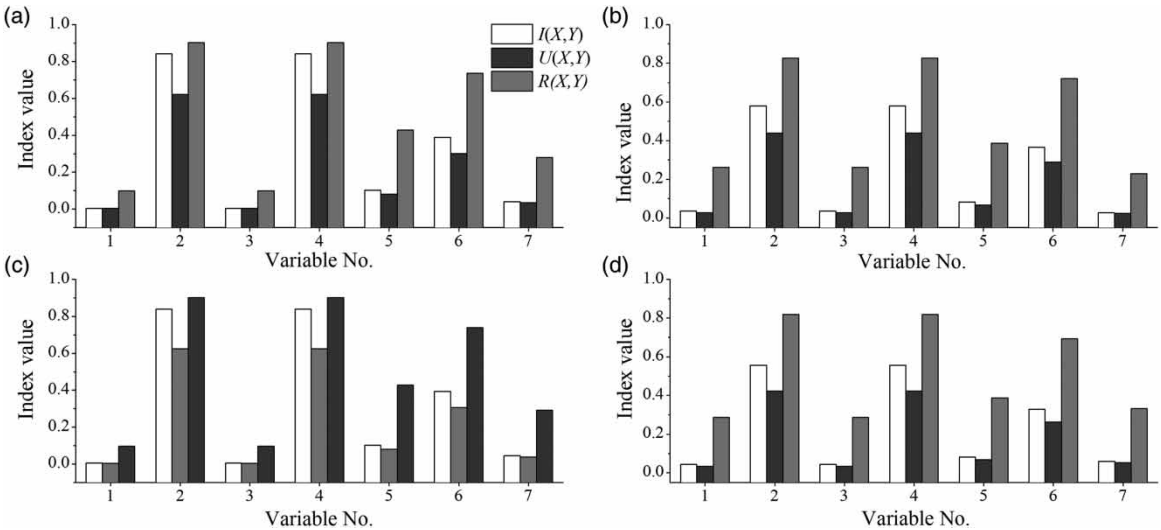

Figure 6 | Index values of input variables in mutual entropy analysis. First and second rows represent input parameters are sampled from uniform and normal distributions, respectively. The plots (a), (c) and the plots (b), (d) indicate output variables (Y) are the mean and variance of GLS, respectively.

GLS. Moreover, when the model input parameters are sampled from uniform and normal distributions, respectively, the influences of input variables are similar for these two distributions.

Compared with the results of stepwise regression analysis, the distance from constant head boundary (D4) has a significant influence on the first two moments of GLS. However, the variable D4 has been excluded from the stepwise regression analyses of both mean and variance of GLS. As shown in Figure 1, the sum of the distances from an observation point to the river boundary and constant head boundary is a constant (the width of the study area, 5,000 m). According to the constructing mechanism of stepwise regression model, the influence of D2 and D4 is presented by only one variable (D2) in stepwise regression analysis. Moreover, the importance of D4 is significant as well as D2. The influence mode of D2 on the output variables is inversed to that of D4, and this situation is the same as D1 and D3. In addition, the relationship between the influence modes of D2 and D4 (or D1 and D3) on output variables is certified by the contingency tables in Tables 4–7. Furthermore, the input variables excluded from the stepwise regression model are able to be identified by mutual entropy analysis. By contrast, these variables are roughly treated as invalid influencing factors by stepwise regression analysis.

Classification tree analysis

Figure 7 displays a conventional diagram that labels the PDF of GLS, when input parameters are sampled from uniform distribution. Figure 8 shows the PDF of GLS when input parameters are sampled from normal distribution.

Figure 3 shows that the PDF of GLS is strongly related to the probability distribution of groundwater model input parameters. However, as shown in Figures 7 and 8, the PDF of GLS is not fully controlled by the probability distribution of input parameters. The category of the PDF of GLS is not uniformly distributed in the space of model layer. For identifying the driving factors that lead GLS to follow the specific PDF (uniform or normal), classification tree method is used to identify these driving factors. The GLS are classified into two categories: 0 obeys uniform or normal distribution when the input parameters are sampled from uniform or normal distribution, respectively; 1 does not obey. The input variables in the classification tree model have identical numbers to the variables used in stepwise regression and mutual entropy analyses (see Table 3).

As shown in Figure 9, GLS is passed into subspaces by selecting suitable input variables used for splitting. In addition, the classification tree (four ranks in this paper) is built by constantly splitting. The maximum purity was set as 0.82 in this classification tree. The results indicate that

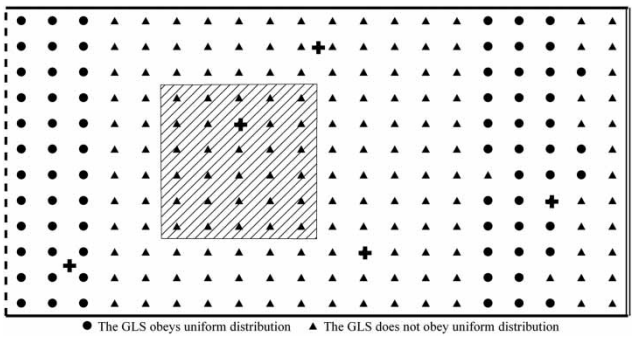

Figure 7 | Conventional diagram labeling the probability distribution of GLS when input parameters are sampled from uniform distribution.

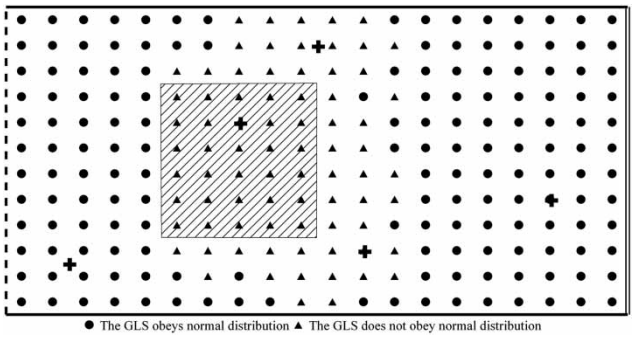

Figure 8 | Conventional diagram labeling the probability distribution of GLS when input parameters are sampled from normal distribution.

when the groundwater model input parameters are sampled from uniform distribution, only two variables (D6 and D2) entered into the classification tree model. Therefore, the probability distribution of GLS is driven by D6 and D2. Furthermore, when the model input parameters are sampled from normal distribution, the tree model contains four variables, and the entry order is D6, D2, D5, and D1. Moreover, variable D6 and D2 are also the most significant driving factors.

Groundwater is a complex system affected by many factors. According to the central limit theorem, when a system is constructed by a large number of independent random variables, each with finite mean and variance, the output of the system will be approximately normally distributed. Thus, when the groundwater model parameters are sampled from normal and uniform distributions, respectively, the outputs of groundwater model following normal distribution are many more than that following uniform distribution (see Figures 3, 4, 7 and 8). Moreover, Figure 9 shows that whether the GLS obey normal distribution is controlled by more driving factors than that leading GLS to obey uniform distribution.

As has been stated, the key driving factors of GLS are D2 and D6. In addition, variable D2 obtains a significant importance in stepwise regression and mutual entropy analyses. However, the mean and variance of GLS are slightly influenced by variable D6. As a result, the mean and variance of GLS are both controlled by the distance from observation point to river boundary (or constant head boundary). The category of the PDF of GLS is dominated by the average distance from observation point to five pumping wells, and the distance from observation point to river boundary (or constant head boundary).

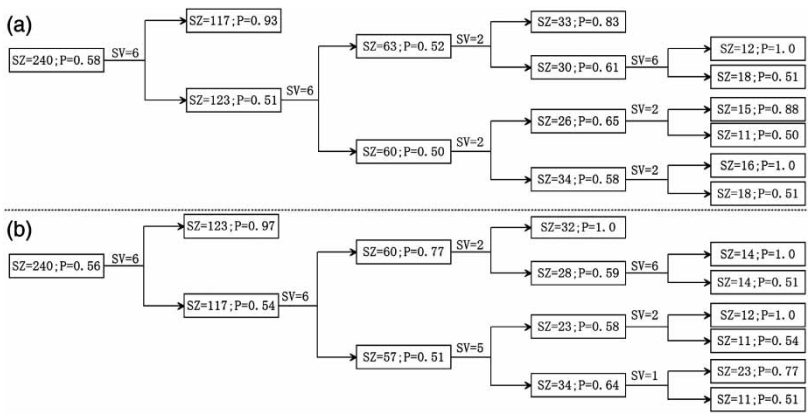

Figure 9 | Splitting process of classification tree analysis. *SZ* denotes sample size, *P* denotes the purity of a space, *SV* denotes splitting variable. (a) and (b) indicate groundwater model parameters are sampled from uniform and normal distribution, respectively.

CONCLUSIONS

The uncertainty of groundwater modeling can be represented by the characteristics of probability distribution of model outputs. Based on a synthetic groundwater model, and the sensitivity analysis of the probability distributions of model outputs, the following conclusions are drawn:

1. The characteristics of probability distribution of groundwater model output is analyzed and summarized. The most important influencing factors for the mean and variance of GLS are the distances from an observation point to river and constant head boundaries. The most important driving factor for the PDF of GLS is the distance from an observation point to all pumping wells. In addition, the distribution characteristics of groundwater model outputs (GLS and budget terms) are significantly influenced by the probability distribution of input parameters.
2. Stepwise regression analysis is a defective sensitivity analysis method for identifying multiple influencing factors which have similar correlation structure with output variable. By contrast, mutual entropy analysis is more general in identifying complicated multivariate relationships. Furthermore, mutual entropy analysis is able to identify the influence of variables which are excluded from stepwise regression analysis. Moreover, classification tree analysis is an effective method for analyzing the key driving factors in a classification output system.

ACKNOWLEDGEMENTS

This study was supported by the National Natural Science Fund of China (Nos. 41172207, 41030746, 51190091, and 41071018), Program for New Century Excellent Talents in University (NCET-12-0262), China Doctoral Program of Higher Education (20120091110026), Qing Lan Project, the Skeleton Young Teachers Program and Excellent Disciplines Leaders in Midlife-Youth Program of Nanjing University.

REFERENCES

Bergante, S., Facciotto, G. & Minotta, G. 2010 Identification of the main site factors and management intensity affecting the establishment of Short-Rotation-Coppices (SRC) in Northern Italy through stepwise regression analysis. *Cent. Eur. J. Biol.* **5**, 522–530.

Blasone, R. S., Vrugt, J. A., Madsen, H., Rosbjerg, D., Robinson, B. A. & Zyvoloski, G. A. 2008 Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov chain Monte Carlo sampling. *Adv. Water Resour.* **31**, 630–648.

Chen, Y. F., Hou, Y., Van Gelder, P. & Zhigui, S. 2002 Study of parameter estimation methods for Pearson-III distribution in flood frequency analysis. *Iahs-Aish P* **271**, 263–269.

Englehart, P. J. & Douglas, A. V. 2010 Diagnosing warm-season rainfall variability in Mexico: A classification tree approach. *Int. J. Climatol.* **30**, 694–704.

Esther, A., Groeneveld, J., Enright, N. J., Miller, B. P., Lamont, B. B., Perry, G. L. W., Blank, F. B. & Jeltsch, F. 2010 Sensitivity of plant functional types to climate change: classification tree analysis of a simulation model. *J. Veg. Sci.* **21**, 447–461.

Gungor, O. & Goncu, S. 2013 Application of the soil and water assessment tool model on the Lower Porsuk Stream Watershed. *Hydrol. Process.* **27**, 453–466.

Haktanir, T. 1992 Comparison of various flood frequency-distributions using annual flood peaks data of rivers in Anatolia. *J. Hydrol.* **136**, 1–31.

Harbaugh, A. W. 2005 The U.S. Geological Survey modular groundwater model–the Ground-Water Flow Process. *U.S. Geological Survey Techniques and Methods 6-A16*, pp. 81–84.

Hashemi, H., Berndtsson, R., Kompani-Zare, M. & Persson, M. 2013 Natural vs. artificial groundwater recharge, quantification through inverse modeling. *Hydrol. Earth Syst. Sci.* **17**, 637–650.

Hassan, A. E., Bekhit, H. M. & Chapman, J. B. 2008 Uncertainty assessment of a stochastic groundwater flow model using GLUE analysis. *J. Hydrol.* **362**, 89–109.

Huysmans, M., Madarasz, T. & Dassargues, A. 2006 Risk assessment of groundwater pollution using sensitivity analysis and a worst-case scenario analysis. *Environ. Geol.* **50**, 180–193.

Katz, R. W., Parlange, M. B. & Naveau, P. 2002 Statistics of extremes in hydrology. *Adv. Water Resour.* **25**, 1287–1304.

Lang, M., Pobanz, K., Renard, B., Renouf, E. & Sauquet, E. 2010 Extrapolation of rating curves by hydraulic modelling, with application to flood frequency analysis. *Hydrol. Sci. J.* **55**, 883–898.

MacQuarrie, C. J. K., Spence, J. R. & Langor, D. W. 2010 Using classification tree analysis to reveal causes of mortality in an insect population. *Agr. Forest Entomol.* **12**, 143–149.

Mazzilli, N., Guinot, V. & Jourde, H. 2010 Sensitivity analysis of two-dimensional steady-state aquifer flow equations. Implications for groundwater flow model calibration and validation. *Adv. Water Resour.* **33**, 905–922.

McMahon, T. A. & Srikanthan, R. 1981 Log Pearson III distribution – Is it applicable to flood frequency-analysis of Australian streams. *J. Hydrol.* **52**, 139–147.

Melo, I., Tomasik, B., Torrieri, G., Vogel, S., Bleicher, M., Korony, S. & Gintner, M. 2009 Kolmogorov–Smirnov test and its use for the identification of fireball fragmentation. *Phys. Rev. C.* **80**, 024904.

Mishra, S., Deeds, N. E. & RamaRao, B. S. 2003 Application of classification trees in the sensitivity analysis of probabilistic model results. *Reliab. Eng. Syst. Safe.* **79**, 123–129.

Mishra, S., Deeds, N. & Ruskauff, G. 2009 Global sensitivity analysis techniques for probabilistic ground water modeling. *Ground Water* **47**, 730–747.

Morway, E. D., Niswonger, R. G., Langevin, C. D., Bailey, R. T. & Healy, R. W. 2013 Modeling variably saturated subsurface solute transport with MODFLOW-UZF and MT3DMS. *Ground Water* **51**, 237–251.

Mpimpas, H., Anagnostopoulos, P. & Ganoulis, J. 2008 Uncertainty of model parameters in stream pollution using fuzzy arithmetic. *J. Hydroinf.* **10**, 189–200.

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A. & Brown, S. D. 2004 An introduction to decision tree modeling. *J. Chemometr.* **18**, 275–285.

Neppel, L., Renard, B., Lang, M., Ayral, P. -A., Coeur, D., Gaume, E., Jacob, N., Payrastre, O., Pobanz, K. & Vinet, F. 2010 Flood frequency analysis using historical data: accounting for random and systematic errors. *Hydrol. Sci. J.* **55**, 192–208.

Onoz, B. & Bayazit, M. 1995 Best-fit distributions of largest available flood samples. *J. Hydrol.* **167**, 195–208.

Pappenberger, F., Beven, K. J., Ratto, M. & Matgen, P. 2008 Multi-method global sensitivity analysis of flood inundation models. *Adv. Water Resour.* **31**, 1–14.

Robin, M. J. L., Gutjahr, A. L., Sudicky, E. A. & Wilson, J. L. 1993 Cross-correlated random-field generation with the direct Fourier-transform method. *Water Resour. Res.* **29**, 2385–2397.

Rojas, R., Feyen, L. & Dassargues, A. 2009 Sensitivity analysis of prior model probabilities and the value of prior knowledge in the assessment of conceptual model uncertainty in groundwater modelling. *Hydrol. Process.* **23**, 1131–1146.

Ross, S. M. 2004 *Introduction to Probability and Statistics for Engineers and Scientists*. Elsevier Academic Press, San Diego, CA.

Singh, V. P. & Singh, K. 1985a Derivation of the gamma-distribution by using the principle of maximum-entropy (POME). *Water Resour. Bull.* **21**, 941–952.

Singh, V. P. & Singh, K. 1985b Derivation of the Pearson Type (PT) III distribution by using the principle of maximum-entropy (POME). *J. Hydrol.* **80**, 197–214.

Smakhtin, V. U. 2001 Low flow hydrology: a review. *J. Hydrol.* **240**, 147–186.

Sun, C. X. & Zheng, S. Q. 2006 Some results of parameter estimator based on uniform distribution. *Coll. Math. J.* **22**, 130–134.

Vogel, R. M., Mcmahon, T. A. & Chiew, F. H. S. 1993 Floodflow frequency model selection in Australia. *J. Hydrol.* **146**, 421–449.

Wang, D., Singh, V. P., Zhu, Y. S. & Wu, J. C. 2009 Stochastic observation error and uncertainty in water quality evaluation. *Adv. Water Resour.* **32**, 1526–1534.

Wang, F. G. & Wang, X. D. 2010 Fast and robust modulation classification via Kolmogorov-Smirnov test. *IEEE T. Commun.* **58**, 2324–2332.

Wu, J. C., Lu, L. & Tang, T. 2011 Bayesian analysis for uncertainty and risk in a groundwater numerical model's predictions. *Hum. Ecol. Risk Assess.* **7**, 1310–1331.

Ye, M., Pohlmann, K. F., Chapman, J. B., Pohll, G. M. & Reeves, D. M. 2010 A model-averaging method for assessing groundwater conceptual model uncertainty. *Ground Water* **48**, 716–728.

Zeng, X. K., Wang, D. & Wu, J. C. 2012 Sensitivity analysis of the probability distribution of groundwater level series based on information entropy. *Stoch. Environ. Res. Risk Assess.* **26**, 345–356.

Zeng, X. K., Wang, D., Wu, J. C. & Chen, X. 2013 Reliability analysis of the groundwater conceptual model. *Hum. Ecol. Risk Assess.* **19**, 515–525.

Zhang, P., Aagaard, P., Nadim, F., Gottschalk, L. & Haarstad, K. 2009 Sensitivity analysis of pesticides contaminating groundwater by applying probability and transport methods. *Integr. Environ. Assess. Manag.* **5**, 414–425.

Zhang, X., Hoermann, G. & Fohrer, N. 2012 Parameter calibration and uncertainty estimation of a simple rainfall-runoff model in two case studies. *J. Hydroinformat.* **14**, 1061–1074.

First received 5 January 2013; accepted in revised form 13 June 2013. Available online 12 July 2013