*J. Earth Syst. Sci.* (2020)129 48
https://doi.org/10.1007/s12040-019-1299-y




# An improved method for predicting water shortage risk in the case of insufficient data and its application in Tianjin, China

Longxia Qian[1,*], Zhengxin Wang[1], Hongrui Wang[2] and Caiyun Deng[2]

[1]*School of Science, Nanjing University of Posts and Telecommunications, Nanjing 210 023, China.*
[2]*College of Water Sciences, Beijing Normal University, Beijing 100 875, China.*
**Corresponding author. e-mail: qianlongxia@njupt.edu.cn*

MS received 8 July 2019; revised 18 September 2019; accepted 19 September 2019

It is very important to estimate the parameters of a risk prediction model in the case of small samples. This paper proposed an improved method for predicting water shortage risk in situations when insufficient data are available. The new method (maximum entropy estimation, MEE) does not require the data about water shortage risk but only a few data about the risk factors. Twelve simulations or experiments were made to evaluate the performance of MEE under different small sample size and compared with the maximum likelihood estimation (MLE) which requires a large amount of data about risk and its factors, and two models which require small samples about risk and risk factors. The result shows that MEE performs much better than MLE, and has an advantage over the two models. Water shortage risks in 2020 in all the districts or counties of Tianjin were predicted by using the new method. The result shows that the values of water shortage risk in most of the districts or counties of Tianjin are very high when the transferred and unconventional water are not used. After using the transferred and unconventional water, all the values of water shortage risk decline considerably.

**Keywords.** Maximum entropy estimation; maximum likelihood estimation; small samples; water shortage risk; Tianjin.

## 1. Introduction

Water scarcity has become an ever-more serious issue and attracted much attention as a result of the rapid expansion of population and economy development, which exerts a negative influence on sustainable development of water resources (Guhathakurta and Saji 2013; Qian *et al.* 2014; Jia *et al.* 2015; Liu and Gan 2018; Yan *et al.* 2018). It is very important to obtain an accurate prediction for water shortage risk in order to take some effective strategies to reduce its potential impacts (Yan *et al.* 2014).

Published online: 09 January 2020

In the past, some statistical methods or models have been widely used for predicting water resources risk. For example, Zhang *et al.* (2016) applied a support vector machine method to evaluate five indicators of water shortage risk, including hazard rate, restorability, vulnerability, and so on. Yu *et al.* (2014) used the Monte Carlo method to forecast water shortage probabilities. Feng and Huang (2008) built a model to assess water shortage risk based on information diffusion theory. Zhang *et al.* (2013) applied a discriminant analysis model to evaluate drought risk. Qian *et al.* (2014) used a linear discriminant analysis

model to predict risk between water supply and demand. Qian *et al.* (2018b) proposed an improved projection pursuit model for risk assessment with high-dimensional indicators. In addition, the logistic regression model was also used for prediction of water quality risk (Yerel and Anagun 2010). However, sometimes a statistical method or model is generally infeasible, because estimating the parameters of a statistical model requires a large amount of observed values about risk and risk factors (Coron *et al.* 2018). Insufficient data is often the case in water resources risk (Tidwell *et al.* 2005). Some models have been proposed for risk assessment in situations when insufficient data are available. For example, Huang (1997) put forward an information diffusion model (IDM) for insufficient data, but the data must follow symmetrical and normal distribution (Bai *et al.* 2015). Then Bai *et al.* (2015) proposed an improved IDM model (IIDM) based on vibrating string equation for extracting more information from sparse data when the data follows asymmetrical and abnormal distribution. However, both models require some data about risk and risk factors. How to assess or predict risk when only small samples about risk factors are available requires further research. As we all know, the MLE is widely applied to estimate the parameters of statistical models, and it has high requirement for data quantity about risk and its factors. Generally, very few qualitative values about risk can be obtained, as water shortage risk is difficult to observe. It means that the MLE is infeasible for parameter estimation in the condition of small samples. How to estimate the parameters of a statistical model in such situation, especially when there are only small samples about risk factors? Qian *et al.* (2018a) proposed an improved model to predict the extreme risk in situations with small samples about the observations. Inspired by this, we want to propose a new method to estimate the parameters of a logistic regression model in situations when only insufficient data about risk factors is available.

Our paper has the following contributions. First, we put forward a new method to estimate the parameters of the logistic regression model in situations with small samples. The new method does not require observed values about water shortage risk, but only some observed values of the risk factors. Then, we used genetic algorithm to search for the optimal parameters.

## 2. Methods

### 2.1 *A new method to estimate the parameters of logistic regression model*

A logistic regression model, one of the risk prediction methods, is often used to study the nonlinear relation between a binary categorical or multi-categorical variable and its factors. Water shortage risk is a binary categorical variable since water shortage happens or does not happen in a certain situation. Therefore, a logistic regression model can be applied to predict water shortage risk. Suppose the risk factors are $\{x_{ij}\,(i=1,2,\ldots,n;j=1,2,\ldots,m)\}$ and water shortage risk samples are $\{y_i\,(i=1,2,\ldots,n)\}$,

$$y_i=\begin{cases}0, & \text{Water shortage risk occurs}\\ 1, & \text{Water shortage risk does not occur}\end{cases}.$$

The number of factors and samples are $m$ and $n$.

$R_i$ is the occurrence probability of water shortage risk with the factors of $x_{ij}\,(i=1,2,\ldots,n;j=1,2,\ldots,m)$. The risk can be predicted as follows (Brown 1982):

$$R_i=\frac{1}{1+e^{-(\alpha+\beta_1x_{i1}+\beta_2x_{i2}+\ldots+\beta_mx_{im})}}, \tag{1}$$

where $\alpha,\beta_1,\beta_2,\ldots,\beta_m$ are the unknown parameters and they are often estimated by MLE. As for MLE, the amount of samples is at least 10 times the number of variables. Therefore, the observed data about risk and its factors with the number of $10m$ is required to compute the parameters. Unfortunately, it is difficult to obtain the observed data about risk. In this case, how can we estimate the parameters of the logistic regression model with only some data about risk factors? To solve the problem, a new method (MEE) will be proposed to estimate the parameters of the logistic regression model with small samples about risk factors. It has the following principle.

For an observation of risk ($R_i$), we can define its entropy to evaluate its degree of uncertainty about the observation. The entropy of $R_i$ is defined as follows (Jones and Jones 2000):

$$\begin{aligned}H(R_i)&=-C[R_i\ln R_i+(1-R_i)\ln(1-R_i)]\\&=-C\left[R_i\ln\left(\frac{R_i}{1-R_i}\right)+\ln(1-R_i)\right]\\&=-C\left\{\frac{\left(\alpha+\sum_{j=1}^{m}\beta_jx_{ij}\right)}{1+\exp\left[-\left(\alpha+\sum_{j=1}^{m}\beta_jx_{ij}\right)\right]}\right.\\&\quad\left.-\ln\left(1+\exp\left(\alpha+\sum_{j=1}^{m}\beta_jx_{ij}\right)\right)\right\},\end{aligned} \tag{2}$$

$C$ is a positive number. According to the principle of maximum entropy (POME), when making the inferences based on sparse data, the parameters to be drawn must have the maximum entropy permitted by the available information from the sparse data (Singh 1997). Moreover, the entropy of an isolated system tends to reach a maximum, and the POME is in conformity with the first principle that the solution should be consistent with the samples with the least hypotheses regarding the unknown parts when insufficient data is available (Jones and Jones 2000). Therefore, the optimal parameters can be obtained when all the values of $H(R_i)\,(i = 1, 2, \ldots, n)$ reaches a maximum (Jones and Jones 2000). Therefore, the optimal parameters can be computed as follows:

$$\max\ H(R_i) = -C\left\{\frac{\left(\alpha + \sum_{j=1}^{m}\beta_j x_{ij}\right)}{1 + \exp\left[-\left(\alpha + \sum_{j=1}^{m}\beta_j x_{ij}\right)\right]} - \ln\left(1 + \exp\left(\alpha + \sum_{j=1}^{m}\beta_j x_{ij}\right)\right)\right\}. \tag{3}$$

However, there are $R_i$ with the number of $n$, and it is impossible to find parameters which maximize all the $H(R_i)\,(i = 1, 2, \ldots, n)$ simultaneously. Therefore, the maximum value of the sequences $(\max\ H(R_i)\,(i = 1, 2, \ldots, n))$ can be chosen as the objective function. To sum up, the model for obtaining the optimal parameters is as follows.

$$\text{maxmax}\, H(R_i) = -C\left\{\frac{\left(\alpha + \sum_{j=1}^{m}\beta_j x_{ij}\right)}{1 + \exp\left[-\left(\alpha + \sum_{j=1}^{m}\beta_j x_{ij}\right)\right]} - \ln\left(1 + \exp\left(\alpha + \sum_{j=1}^{m}\beta_j x_{ij}\right)\right)\right\}, \tag{4}$$

Equation (4) does not need samples of water shortage risk, and only risk factors are required.

### 2.2 *Searching the optimal parameters through genetic algorithm*

Genetic algorithm is a method of searching the optimal solution by modelling natural evolution process, and it can obtain global optima (Goldberg and Holland 1988). Genetic algorithm has the following steps. First, initial population is generated based on some rules. Second, the individual fitness is computed to evaluate the quality of the individual. The value of the objective function (equation 4) can be used as the fitness. Individuals with higher fitness will have more chances to inherit to the next generation. Then, new generations will be produced by selection operation, crossover operation and mutation operation. More details about the computation process can be found in the literature (Goldberg and Holland 1988).

## 3. Model validation

### 3.1 *Comparison with the MLE*

Random number generator of logistic regression was applied to generate a sequence with six parameters and its sample size is 1000. Ten simulations were made to test the performance of MEE under different small sample sizes (100, 90, 80, 70, 60, 50, 40, 30, 20 and 10). The values of the error generated from MLE and MEE are shown in tables 1 and 2. The error is calculated as follows.

$$\text{Error} = \frac{|p_i - P|}{P}, \tag{5}$$

where $p_i$ is the parameter value generated from MEE under sample sizes of 100, 90, 80, 70, 60, 50, 40, 30, 20 and 10, respectively, and $P$ is the MLE-estimated parameter under sample size of 1000.

Tables 1 and 2 show that MEE performs much better than MLE. For example, MEE can obtain a satisfactory result when there are only 20 samples, while MLE's poor performance can be declared

Table 1. *The APE values generated from MLE under different sample sizes.*

| APE | 100 | 90 | 80 | 70 | 60 | 50 | 40 | 30 | 20 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | 0.8% | 21.0% | 198% | 132100% | 132100% | 131800% | 131800% | 131800% | 134000% | / |
| $\beta_1$ | 62.4% | 49.2% | 86.7% | 84900% | 85500% | 85500% | 85500% | 85500% | 85500% | / |
| $\beta_2$ | 1.3% | 50.1% | 83.3% | 260400% | 260800% | 260800% | 260800% | 260800% | 268600% | / |
| $\beta_3$ | 34.8% | 73.9% | 8.7% | 214900% | 215200% | 215200% | 215200% | 215200% | 220700% | / |
| $\beta_4$ | 18100% | 20100% | 16100% | 30306100% | 30332100% | 30332100% | 30332100% | 30332100% | 31026100% | / |
| $\beta_5$ | 88.4% | 27.3% | 335% | 144900% | 144800% | 144800% | 144800% | 144800% | 153200% | / |

Table 2. *The APE values generated from MEE under different sample sizes.*

| APE | 100 | 90 | 80 | 70 | 60 | 50 | 40 | 30 | 20 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| $\alpha$ | 2.3% | 2.5% | 3.6% | 3.6% | 3.6% | 3.6% | 3.6% | 3.6% | 3.6% | 43.1% |
| $\beta_1$ | 0.6% | 0.1% | 0.2% | 0.2% | 0.2% | 0.2% | 0.2% | 0.2% | 0.2% | 21.2% |
| $\beta_2$ | 0.9% | 0.7% | 0.9% | 0.9% | 0.9% | 0.9% | 0.9% | 0.9% | 0.9% | 26.2% |
| $\beta_3$ | 10.4% | 14.3% | 11.6% | 11.6% | 11.6% | 11.6% | 11.6% | 11.6% | 11.6% | 11.6% |
| $\beta_4$ | 0% | 45.0% | 55.0% | 55.0% | 55.0% | 55.0% | 55.0% | 55.0% | 55.0% | 55.0% |
| $\beta_5$ | 2.3% | 0.7% | 0.5% | 0.5% | 0.5% | 0.5% | 0.5% | 0.5% | 0.5% | 18.8% |

Table 3. *Samples of marine environmental risk for naval activity (Qian et al. 2018b).*

| Sl. no. | Wind speed (m/s) | Wave height (m) | Visibility (km) | Thunder possibility | Low cloud | Risk | Probability |
|---|---|---|---|---|---|---|---|
| 1 | 0 | 0.1 | 10 | 0 | 0.1 | 0 | 0 |
| 2 | 38 | 14 | 1 | 0.5 | 0.6 | 1 | 1 |
| 3 | 3 | 0.1 | 0.5 | 0.1 | 0 | 0.3 | 0 |
| 4 | 3 | 0.2 | 8 | 0 | 0.3 | 0.1 | 0 |
| 5 | 8 | 2 | 6 | 0.4 | 0.1 | 0.4 | 0 |
| 6 | 19 | 6 | 8 | 0.5 | 0.2 | 0.7 | 1 |
| 7 | 5 | 0.2 | 8 | 0.7 | 0.4 | 0.5 | 1 |
| 8 | 2 | 0.1 | 3 | 0.2 | 0 | 0.2 | 0 |
| 9 | 1 | 0.1 | 9 | 0.2 | 0.7 | 0.3 | 0 |
| 10 | 4 | 0.3 | 10 | 0.1 | 0 | 0.1 | 0 |
| 11 | 22 | 8 | 4 | 0.4 | 0.2 | 0.6 | 1 |
| 12 | 2 | 0 | 0.4 | 0 | 0.1 | 0.4 | 0 |
| 13 | 8 | 2 | 8 | 0.1 | 0.5 | 0.2 | 0 |
| 14 | 16 | 3 | 5 | 0.8 | 0.6 | 0.8 | 1 |
| 15 | 10 | 3 | 3 | 0.1 | 0.3 | 0.3 | 0 |
| 16 | 2 | 0.2 | 10 | 0 | 0.1 | 0 | 0 |
| 17 | 18 | 5 | 3 | 0.8 | 0.6 | 0.9 | 1 |
| 18 | 17 | 5 | 4 | 0.3 | 0.3 | 0.4 | 0 |
| 19 | 5 | 1 | 5 | 0.4 | 0.9 | 0.6 | 1 |
| 20 | 4 | 0.2 | 2 | 0 | 0.1 | 0.2 | 0 |
| 21 | 3 | 0.1 | 9 | 0.4 | 0.5 | 0.3 | 0 |
| 22 | 1 | 0.1 | 0.4 | 0.3 | 0.6 | 0.6 | 1 |
| 23 | 20 | 6 | 1 | 0.9 | 0.7 | 1 | 1 |
| 24 | 12 | 3 | 7 | 0.3 | 0 | 0.3 | 0 |
| 25 | 8 | 1.5 | 6 | 0.1 | 0.4 | 0.2 | 0 |
| 26 | 4 | 0.2 | 0.2 | 0.2 | 0.8 | 0.8 | 1 |
| 27 | 6 | 1 | 8 | 0.4 | 0.6 | 0.4 | 0 |
| 28 | 15 | 4 | 5 | 0.3 | 0.3 | 0.5 | 1 |
| 29 | 10 | 3 | 6 | 0.2 | 0.5 | 0.4 | 0 |
| 30 | 3 | 0.1 | 0.3 | 0.4 | 0.7 | 0.7 | 1 |
| 31 | 5 | 1 | 6 | 0.5 | 0.9 | 0.6 | 1 |
| 32 | 6 | 0.5 | 7 | 0 | 0 | 0.1 | 0 |
| 33 | 0 | 0 | 8 | 0.1 | 0.6 | 0.2 | 0 |
| 34 | 3 | 0.1 | 0.5 | 0.4 | 0.2 | 0.5 | 1 |
| 35 | 2 | 0.1 | 3 | 0.2 | 0.9 | 0.5 | 1 |
| 36 | 2 | 0.1 | 1 | 0.8 | 0.5 | 0.8 | 1 |

below 70 samples. Moreover, MEE provides an acceptable result with only 10 samples, while MLE is inapplicable when only 10 samples are used.

### 3.2 *Comparison with other models*

In this section, two experiments are performed to compare the logistic regression model with MEE

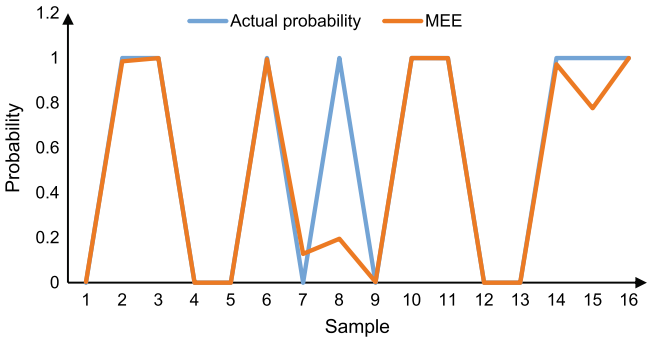


Figure 1. Comparison of the actual and predicted probabilities generated from the MEE using 20 training samples.

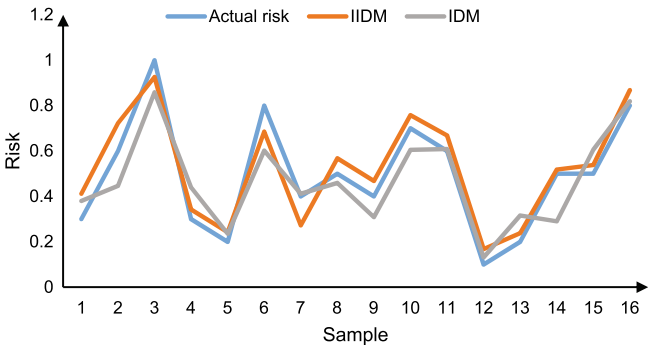


Figure 2. Comparison of the actual and predicted risks generated from the IDM and IIDM using 20 training samples.

Table 4. *Contrast of the average error between three models.*

| | Model | | |
|---|---|---|---|
| Error | MEE | IDM | IIDM |
| The average error | 0.076 | 0.092 | 0.071 |

with the models mentioned in the Introduction (IDM and IIDM). The data reported by Qian *et al.* (2018b) was used for comparison, as shown in table 3. There are some differences between the three models, although all of them are used for sparse data. The logistic regression model with MEE is applied to predict probability, while the IDM and IIDM belong to fitting models. Therefore, the results generated from the logistic regression model with MEE should be compared with the probabilities (the last column of table 3), and those generated from the IDM and IIDM should be compared with the risks (the seventh column of table 3). Twenty samples in table 3 are used for model construction, and the remaining 16 samples are used for model validation. The values of probability or risk generated from the three models are shown in figures 1 and 2. The average errors between the actual and predicted probabilities or risks are also computed for comparison, as shown in table 4. Figure 1 indicates a good match between the actual and predicted probabilities generated by the logistic regression model with MEE and there only exist deviations in the 7th, 8th, and 15th samples. Figure 2 shows that the IIDM clearly outperforms the IDM. Table 4 shows that both the MEE and IIDM obtain a satisfactory result for the average error. Overall, the IIDM performs best, followed by the MEE. The performance of the MEE is very satisfying considering that it does not require the data about risk but only

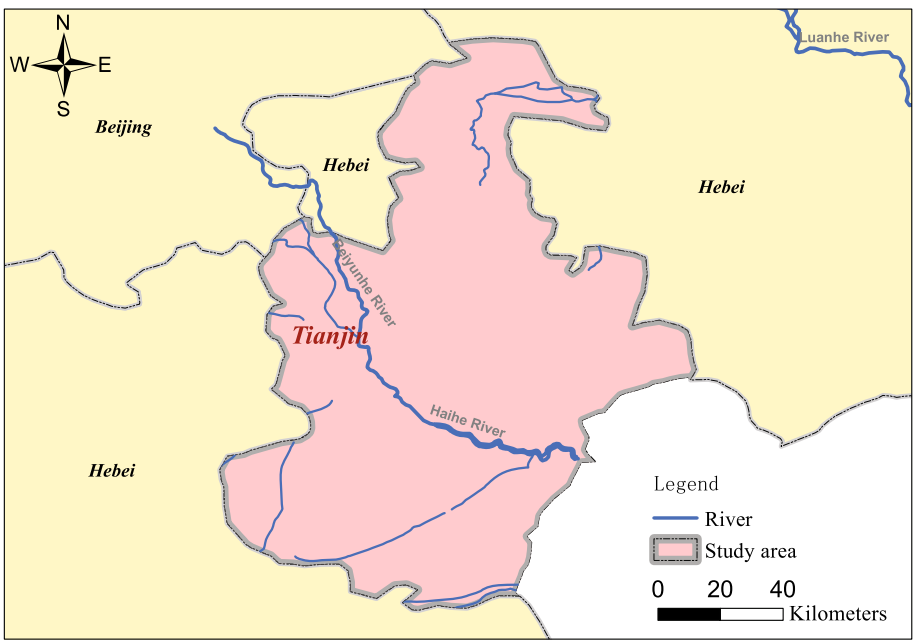


Figure 3. Distribution of river system of Tianjin.

Table 5. *The values of water shortage risk factors of Tianjin.*

| Region | Water shortage rate (%) | Water shortage per capita ($m^3$ per capita) | Groundwater utilization rate (%) | Water consumption per GDP ($m^3$ per 10,000 CNY) |
|---|---|---|---|---|
| Six districts of the city | 91.3 | 61 | 173.5 | 12 |
| New district of Binhe | 31.3 | 214 | 176.9 | 29 |
| Dongli district | 12.9 | 107 | 83.3 | 31 |
| Xiqing district | 40.5 | 245 | 99.5 | 38 |
| Jinnan district | 13.3 | 75 | 211.1 | 43 |
| Beichen district | 32.4 | 154 | 82.9 | 39 |
| Wuqing district | 40.4 | 282 | 185.3 | 218 |
| Baodi district | 33.5 | 282 | 134.6 | 275 |
| Jinghai county | 3.1 | 162 | 99.8 | 133 |
| Ji county | 13.5 | 144 | 225.4 | 189 |

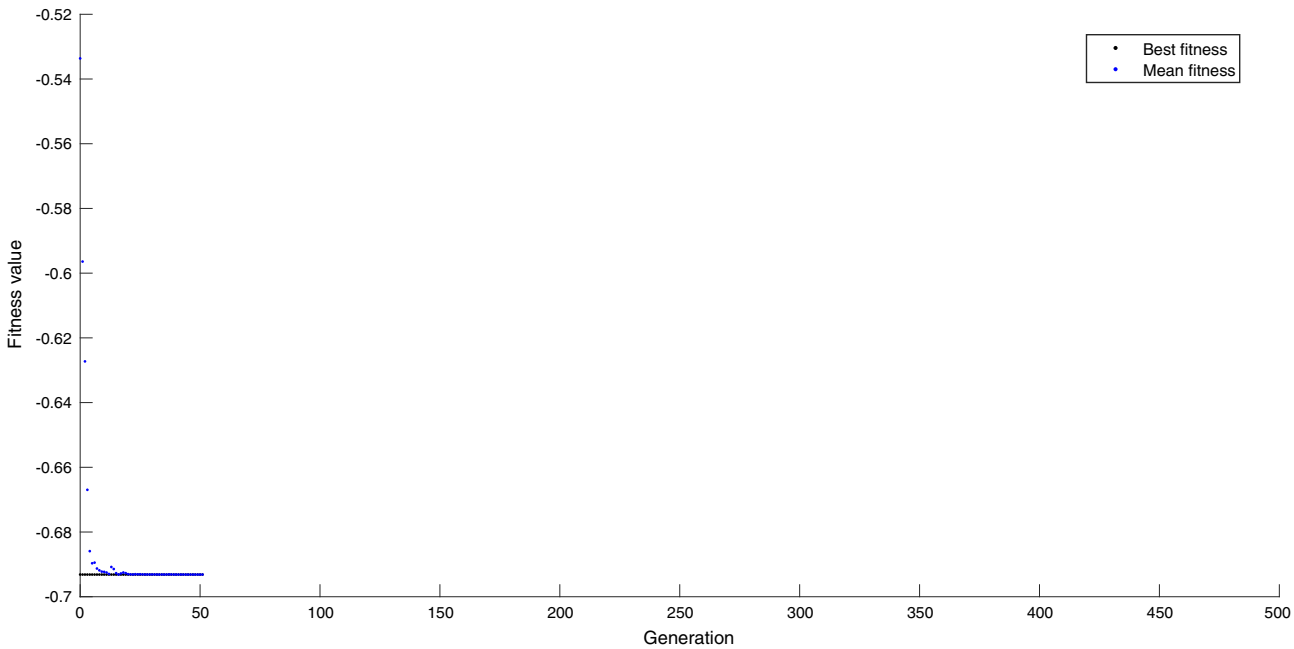


Figure 4. The best fitness value.

a few data about the risk factors, while the IDM and IIDM require some data about risk and risk factors. To sum up, the MEE is reliable in situations with only small samples about risk factors, while the IDM and IIDM are invalid.

## 4. Model application

### 4.1 *Study area*

Tianjin is located in the downstream of Haihe River Basin (figure 3). The average annual precipitation is about 574.9 mm. The average annual amount of surface water, groundwater and total water resources are 10.65, 5.90 and 15.69 billion $m^3$, respectively. The water resources per capita is only 160 $m^3$. Water shortage is very serious, and urban water supply mainly depends on external water sources, such as Luanhe River and South-to-North transferred water. In addition, other unconventional sources, e.g., reclaimed water and seawater desalination are also used for coping with water scarcity. In 2016, 3.43 billion $m^3$ water was drawn from unconventional sources.

Ten samples about water shortage risk factors for Tianjin in 2008 are shown in table 5 (Zheng *et al.* 2011).

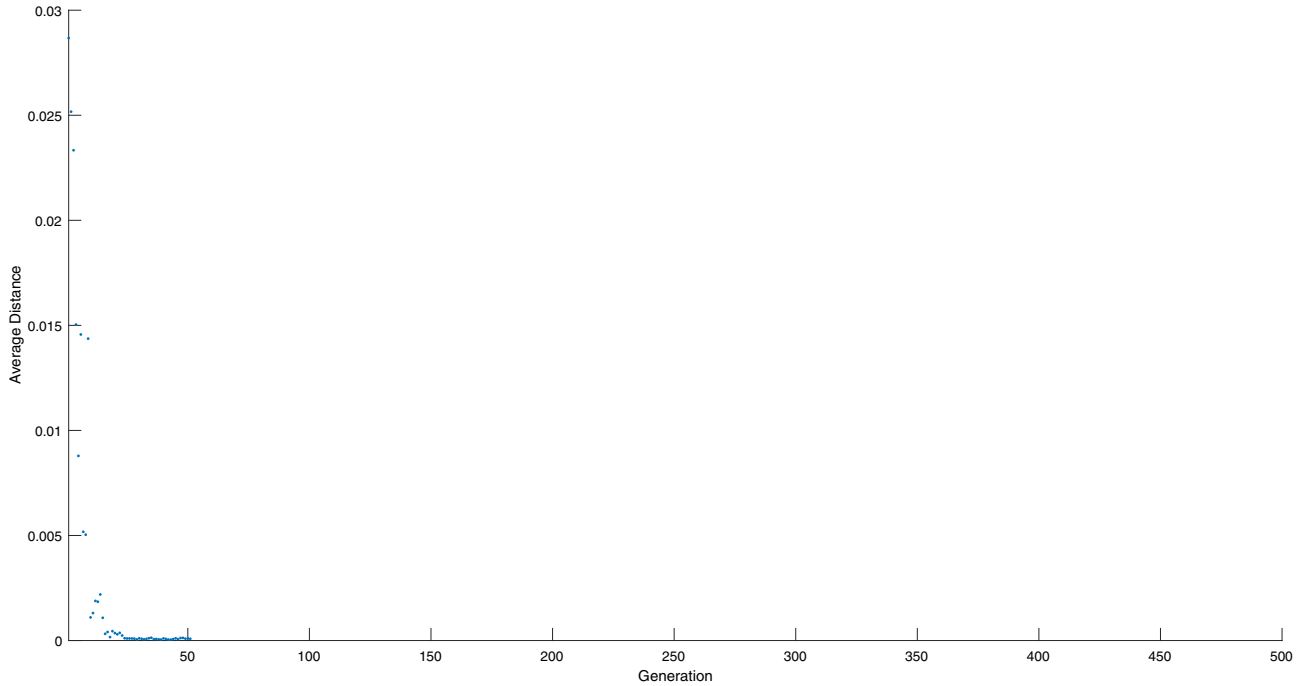


Figure 5. The average distance between individuals.

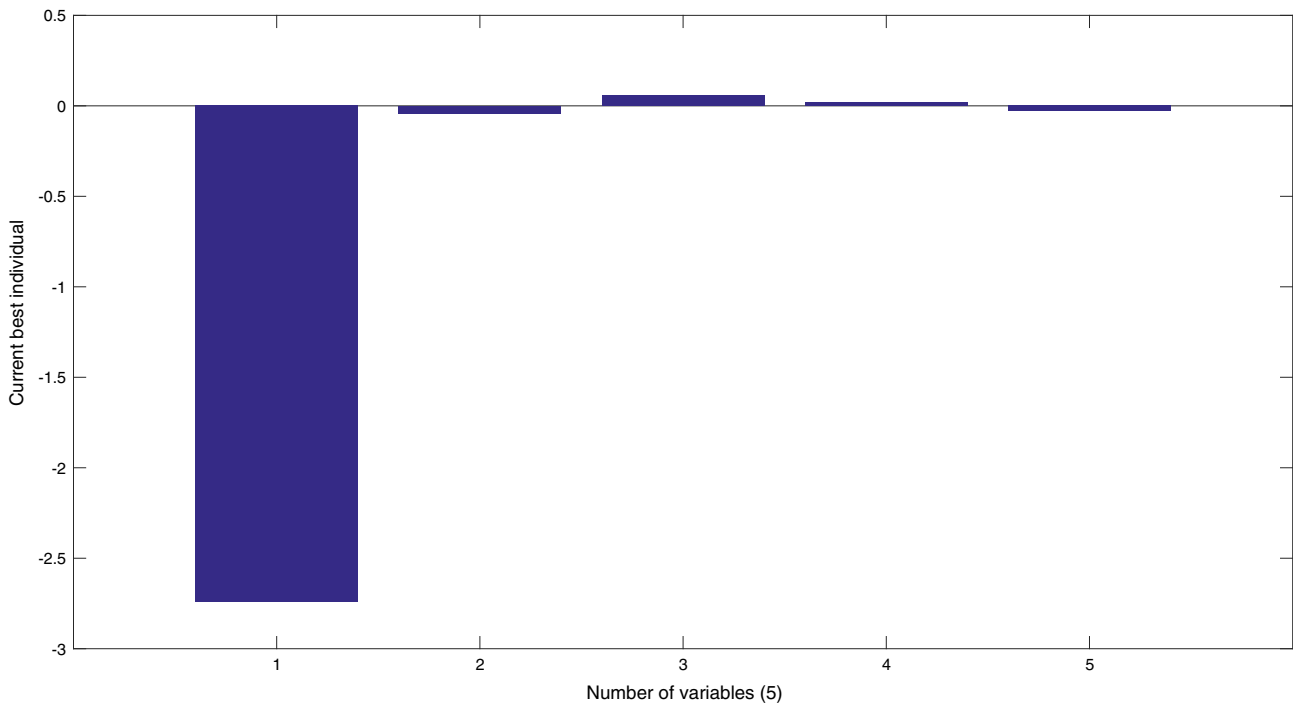


Figure 6. The current best individual.

## 4.2 *Results and analysis*

### 4.2.1 *Parameter estimation by MEE*

Substituting the values of water shortage rate, water shortage per capita, and groundwater utilization rate and water consumption per GDP (table 5) into equation (4), the optimal parameters can be obtained by genetic algorithm. The process of parameter estimation is shown in figures 4–6.

Figures 4 and 5 show that the optimal parameters are obtained in the 51st generation, and the average distance between individuals converges

Table 6. *The amount of water supply and use in 2020 in all the districts or counties of Tianjin (billion $m^3$).*

| District/County | Water supply excluding transferred and unconventional water | Total water supply | Water use |
|---|---|---|---|
| Six districts of the city | 0.38 | 6.46 | 6.46 |
| New district of Binhe | 0.85 | 10.84 | 10.84 |
| Dongli district | 0.02 | 0.37 | 0.37 |
| Xiqing district | 0.09 | 1.56 | 1.56 |
| Jinnan district | 0.05 | 0.8 | 0.80 |
| Beichen district | 0.03 | 0.54 | 0.54 |
| Wuqing district | 0.37 | 0.91 | 0.96 |
| Baodi district | 0.18 | 0.71 | 0.75 |
| Ninghe county | 0.14 | 0.74 | 0.77 |
| Jinghai county | 0.15 | 0.74 | 0.81 |
| Ji county | 0.65 | 0.89 | 0.97 |

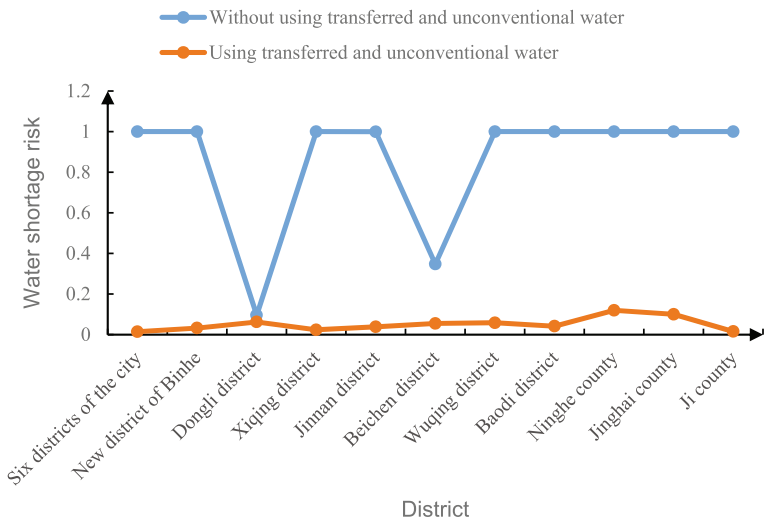


Figure 7. The risk in 2020 in all the district or counties of Tianjin.

to 0. The optimal parameters of $\alpha, \beta_1$ $(j = 1, 2, 3, 4)$ are –2.347, –0.074, 0.084, 0.043, –0.089, respectively (figure 6). Therefore, the water shortage risk prediction model is as follows.

$$R = \frac{1}{1 + e^{(2.347+0.074x_1-0.084x_2-0.043x_3+0.089x_4)}}. \quad (6)$$

### 4.2.2 *Water shortage risk prediction in 2020 in Tianjin*

The amount of water supply and water use in 2020 in all the districts or counties of Tianjin can be found in Zheng *et al.* (2011) as shown in table 6. The population and GDP in 2020 in all the districts or counties of Tianjin are predicted by a grey prediction model, and more details can be found in Qian *et al.* (2016).

According to the amount of water supply and water use, the population and GDP in 2020, the values of water shortage rate, water shortage per capita and water consumption per GDP in 2020 in all the districts or counties of Tianjin without and with using transferred and unconventional water can be obtained. The groundwater utilization rate is assumed to be taken the current value. Substituting the values of water shortage rate, water shortage per capita, and groundwater utilization rate and water consumption per GDP without and with using transferred and unconventional water into equation (6), the water shortage probabilities in 2020 in all the district or counties of Tianjin are shown in figure 7.

Figure 7 shows that the values of water shortage probability in most of the districts or counties are

close to 1 when the transferred and unconventional water are not used. The water shortage probabilities in the districts of Dongli and Beichen are only 0.09 and 0.35. After using the transferred and unconventional water, all the values of water shortage probability decline significantly, by 88.6%. Therefore, the transferred and unconventional water use are of great importance to reduce the water shortage risk.

According to Yan *et al.* (2007), the synchronous encounter probability of poor precipitation in the water source area and Haihe River Basin during dry season is about 0.19. If poor precipitation events happen simultaneously in both water source area and Haihe River Basin, the amount of transferred water would be insufficient to satisfy the water supply. Therefore, the utilization of the unconventional water should be increased in the future.

## 5. Conclusions

Our paper proposed an improved method for predicting water shortage risk in situations when insufficient data are available. The new method (MEE) does not require the data about water shortage risk but only a few data about the risk factors. Twelve simulations or experiments were made to evaluate the performance of MEE under different small sample sizes and compared with the maximum likelihood estimation (MLE) which requires a large amount of data about risk and risk factors, and two models which require small samples about risk and risk factors. The result shows that MEE performs much better than MLE. Although the performance of the MEE is nearly identical to the two models, the MEE has an advantage that it require only a few data about risk factors.

Water shortage risks in 2020 in all the districts and counties of Tianjin were predicted by using the new method. The result shows that the values of water shortage probability in most of the districts or counties are very high when the transferred and unconventional water are not used. After using the transferred and unconventional water, all the values of water shortage probability decline significantly, by 88.6%.

However, there also exist some problems about the MEE. Initial conditions or the range of the parameters need to be given when searching the optimal parameter through genetic algorithm. But sometimes it is very difficult to obtain some information about the initial conditions or the range of the parameters. How to search the optimal parameters will be the focus of our further study.

## Acknowledgement

The study was supported by National Natural Science Foundation of China (Grant No. 51609254).

## References

Bai C Z, Zhang R, Hong M, Qian L and Wang Z 2015 A new information diffusion modelling technique based on vibrating string equation and its application in natural disaster risk assessment; *Int. J. Gen. Syst.* **44(5)** 601–614.

Brown C C 1982 On a goodness-of-fit test for the logistic model based on score statistics; *Commun. Stat.-Theo. Meth.* **11(10)** 1087–1105.

Coron C, Calenge C, Giraud C and Julliard R 2018 Bayesian estimation of species relative abundances and habit preferences using opportunistic data; *Environ. Ecol. Stat.* **25(1)** 71–93.

Feng L H and Huang C F 2008 A risk assessment model of water shortage based on information diffusion technology and its application in analyzing carrying capacity of water resources; *Water Resour. Manag.* **22** 621.

Goldberg D E and Holland J H 1988 Genetic algorithms and machine learning; *Mach. Learn.* **2** 95–99.

Guhathakurta P and Saji E 2013 Detecting changes in rainfall pattern and seasonality index vis-à-vis increasing water scarcity in Maharashtra; *J. Earth Syst. Sci.* **122(3)** 639–649.

Huang C F 1997 Principle of information diffusion; *Fuzzy Sets Syst.* **91(1)** 69–90.

Jia X L, Li C H, Cai Y P, Wang X and Sun L 2015 An improved method for integrated water security assessment in the Yellow River basin, China; *Stoch. Environ. Res. Risk Assess.* **29(8)** 2213–2227.

Jones G A and Jones J M 2000 *Information and coding theory*; Springer-Verlag London Ltd., London.

Liu B and Gan H 2018 Evapotranspiration management based on the application of SWAT for balancing water consumption: A case study in Guantao, China; *J. Earth Syst. Sci.* **127** 51.

Qian L, Wang H and Zhang K 2014 Evaluation criteria and model for risk between water supply and water demand and its application in Beijing; *Water Resour. Manag.* **28** 4433–4447.

Qian L, Zhang R, Hong M, Wang H and Yang L 2016 A new multiple integral model for water shortage risk assessment and its application in Beijing, China; *Nat. Hazards* **80(1)** 43–67.

Qian L, Wang H, Dang S, Wang C, Jiao Z and Zhao Y 2018a Modelling bivariate extreme precipitation distribution for data scarce regions using Gumbel–Hougaard copula with maximum entropy estimation; *Hydrol. Process.* **32** 212–227.

Qian L, Zhang R, Hou T and Wang H 2018b A new nonlinear risk assessment modeling technique based on an improved project pursuit; *Stoch. Environ. Res. Risk Assess.* **32(6)** 1465–1478.

Singh V P 1997 The use of entropy in hydrological and water resources; *Hydrol. Process.* **11** 587–626.

Tidwell V C, Cooper J A and Silva C J 2005 Threat assessment of water supply systems using Markov latent effects modeling; *J. Water Resour. Plann. Manag.* **131(3)** 218–227.

Yan B W, Guo S L and Xiao Y 2007 Synchronous-asynchronous encounter probability of rich-poor precipitation between source area and water receiving areas in the Middle Route of South-North Water Transfer Project; *J. Hydraul. Eng.* **38(10)** 1178–1185 (in Chinese).

Yan D, Weng B and Wang G *et al.* 2014 Theoretical framework of generalized watershed drought risk evaluation and adaptive strategy based on water resources system; *Nat. Hazards* **73(2)** 259–276.

Yan D, Yao M and Ludwig F *et al.* 2018 Exploring future water shortage for large river basins under different water allocation strategies; *Water Resour. Manag.* **32(9)** 3071–3086.

Yerel S and Anagun A S 2010 Assessment of water quality observation stations using cluster analysis and ordinal logistic regression technique; *Int. J. Environ. Pollut.* **42(4)** 344–358.

Yu P S, Yang T C, Kuo C M and Wang Y T 2014 A stochastic approach for seasonal water-shortage probability forecasting based on seasonal weather outlook; *Water Resour. Manag.* **28(12)** 3905–3920.

Zhang Q, Liang X and Fang Z *et al.* 2016 Urban water resources allocation and shortage risk mapping with support vector machine method; *Nat. Hazards* **81** 1209–1228.

Zhang Q, Zhang J, Yan D and Bao Y 2013 Dynamic risk prediction based on discriminant analysis for maize drought disaster; *Nat. Hazards* **65** 1275–1284.

Zheng J, Wu W and Hu X *et al.* 2011 *Integrated risk governance-comprehensive energy and water resources risk in China*; Science Press, Beijing (in Chinese).

Corresponding editor: Subimal Ghosh