An improved logistic probability prediction model for water shortage risk in situations with insufficient data

Longxia Qian[1], Ren Zhang[1,2*], Chengzu Bai[1], Yangjun Wang[1] and Hongrui Wang[3]

[1] Institute of Meteorology and Oceanography, National University of Defense Technology, Nanjing, China, 211101

[2] Collaborative Innovation Center on Forecast Meteorological Disaster Warning and Assessment, Nanjing University of Information Science & Technology, Nanjing, China, 210044

[3] College of Water Sciences, Beijing Normal University, Key Laboratory for Water and Sediment Sciences, Ministry of Education, Beijing , China, 100875

Abstract. In drought years, it is important to have an estimate or prediction of the probability that a water shortage risk will occur to enable risk mitigation. This study developed an improved logistic probability prediction model for water shortage risk in situations when there is insufficient data. First, information flow was applied to select water shortage risk factors. Then, the logistic regression model was used to describe the relation between water shortage risk and its factors, and an alternative method of parameter estimation (maximum entropy estimation) was proposed in situations where insufficient data was available. Water shortage risk probabilities in Beijing were predicted under different inflow scenarios by using the model. There were two main findings of the study. (1) The water shortage risk probability was predicted to be very high in 2020, although this was not the case in some high inflow conditions. (2) After using the transferred and reclaimed water, the water shortage risk probability

* Correspondence to: Ren Zhang, Institute of Meteorology and Oceanography, National University of Defense Technology, Nanjing, China, 211101

E-mail: zrpaper@163.com

declined under all inflow conditions (59.1% on average), but the water shortage risk probability was still high in some low inflow conditions.

Keywords Information flow · Risk factors · Logistic regression model · Maximum entropy estimation · Insufficient data

1 Introduction

Nowadays, water shortages have become a serious problem in many parts of the world due to climate change, heightened demand of water and integrated urbanization, and there is a negative impact on the security and sustainable development of water resources (Giacomelli et al., 2008; Weng et al., 2015; Christodoulou 2011; Wang et al. 2012; Yang et al. 2015 Qian et al. 2014; Li et al. 2014). Risk is a measure of the probability and severity of adverse effects (Haimes, 2009). It is important to have an estimate or prediction of the probability that a water shortage risk will occur so that effective measures for risk mitigation can be developed, particularly in the case of precipitation deficits (drought).

Hashimoto et al. (1982) stated that risk can be described by the probability that a system is in an unsatisfactory state. How to predict or estimate risk probability is still an open issue with no definite solution. Mackenzie (2014) believed that an analyst should first develop a probability distribution over the range of consequences that fully describe the risk of an event. The simulation of probability distribution should be based on a large number of data (Bedford and Cooke, 2001; Giannikopoulou et al., 2015). Unfortunately, a full probabilistic assessment is generally not feasible, because there is insufficient data to quantify the associated probabilities (Tidwell et al., 2005). In some cases, frequency is often used as a substitute for probability in the risk assessment of water resources (Hashimoto et al., 1982; Rajagopalan et al., 2009; Sandoval-Solis et al., 2011), while in other cases, interval-valued probabilities and fuzzy probabilities have been proposed to elaborate the concept of an imprecise probability (Karimi and Hüllermeier, 2007). However, these approaches only consider the probability of the hazard without consideration of the impact of risk factors. The risk factors include characteristics of hazards and existing conditions of vulnerability that could potentially harm exposed people, property, services and so on (UNISDR, 2009). There are many aspects of vulnerability arising from various physical, social, economic, and environmental factors (Qian et al., 2016; Haimes, 2006; UNISDR, 2009). Therefore, it has been concluded that modeling risk probability requires a consideration of vulnerability (Haimes, 2006). Although increasing attention has been given to vulnerability assessment (Villagrán, 2006; Plummer, 2012), there have been few studies of the relation between risk probability and water resources vulnerability.

A water shortage can either occurs or not occur, and therefore water shortage risk is a binary categorical variable. According to statistical theory, a logistic regression model is a nonlinear regression method of studying a binary categorical or multi-categorical variable and its impact factors (Breslow, 1988). Therefore, a logistic regression model can be used to describe the relation between water shortage risk and its impact factors. The parameters of a logistic regression model are often estimated by a maximum likelihood estimation; a large number of observed values of risk (i. e., samples that water shortage risk does or does not occur) and risk factors are required for parameter estimation (Balakrishnan, 1992). However, the statistical data about risk and its factors are insufficient in China. Therefore, the method of maximum likelihood estimation is not applicable when the sample size is small. For this reason, we propose an alternative method of parameter estimation for a logistic regression model when data is insufficient. Moreover, the backward mode is often applied for the selection of sensitive factors, but the calculation is very complicated.

The contributions of our paper are as follows. First, we used a logistic regression model to explore the nonlinear relation between water shortage risk and its factors. Then, we introduced an information flow (Liang, 2014) for the selection of significant risk factors. Compared with the backward mode, it was very easy to determine whether there was a cause and effect between the water shortage risk and its factors. Finally, we proposed an alternative method of parameter estimation (maximum entropy estimation) for a logistic regression model in situations with a lack of data. The new method requires only a few data, while maximum likelihood estimation requires a large amount of data.

The remainder of the paper is organized as follows. Section 2 presents the principles and structure of the logistic probability prediction model for water shortage risk. Section 3 presents the application of the model and the results of the research and Section 4 presents some conclusions and proposes future work.

2 Materials and methods

2.1 Study area

Beijing, China's capital, is located in the northwest of the North China Plain, and consists of five water systems from the east to the west (Figure 1). The average annual precipitation is 585 mm. Precipitation in summer accounts for 70% of the total for the whole year. Beijing, with a population of more than 20 million, is faced with a severe shortage of water resources. The amount of self-generated water resources is only 37.39×10^{8} m^{3}. The amount of water resources per capita is about 200 m^{3}, which is about one eighth of the value of water resources per capita for China and one thirtieth of the global value of water resources per capita.

The available surface water and groundwater is unable to meet the needs of the city's economic and social development. Some measures, such as the use of transferred and reclaimed water have been put in place to mitigate the water shortage. In 2014, through the South-to-North Water Diversion Project, water was channeled from the Danjiangkou Reservoir in central China’s Hebei province to Beijing. Reclaimed water is also essential for Beijing and is mainly used for agricultural irrigation and toilet flushing.

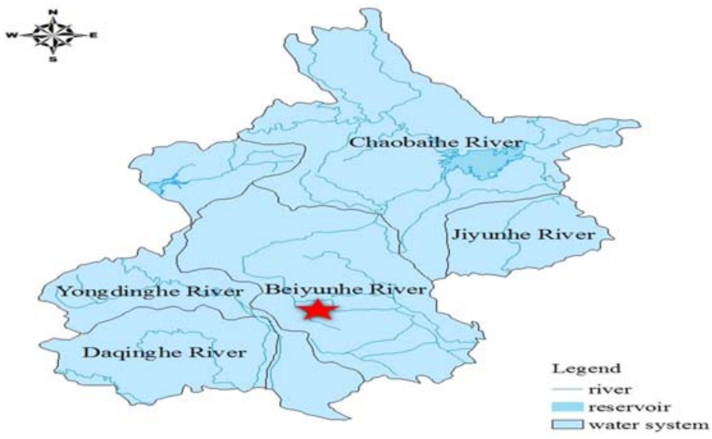

Figure 1. Distribution of water system of Beijing

2.2 Data collection

The data used in this paper were obtained from various sources. The inflow and precipitation sequences from 1956 to 2012 were provided by Beijing Hydrological Station. The water demand for 2020 was based on the Beijing City National Comprehensive Plan for Water Resources (Beijing Municipal Development and Reform Commission and Beijing Municipal Bureau of Water Affairs, 2009). The water supply sequence for 2020 in the inflow conditions of 1956–2012 was computed by an analysis of the balance between water supply and water demand. The population size and gross domestic product (GDP) from 1979 to 2012 were taken from the Statistical Yearbook 2014 of Beijing City (Statistical Bureau of Beijing City, 2014). The total amount of water resources from 1979 to 2012 were provided by Beijing Hydrological Station. The water use statistics and data regarding the treatment of domestic sewage from 1979 to 2012 were taken from the Statistical Yearbook 2014 of Beijing City (Statistical Bureau of Beijing City, 2014).

2.3 Model development

A flowchart showing the operation of the probability prediction model for water shortage risk is given in Figure 2.

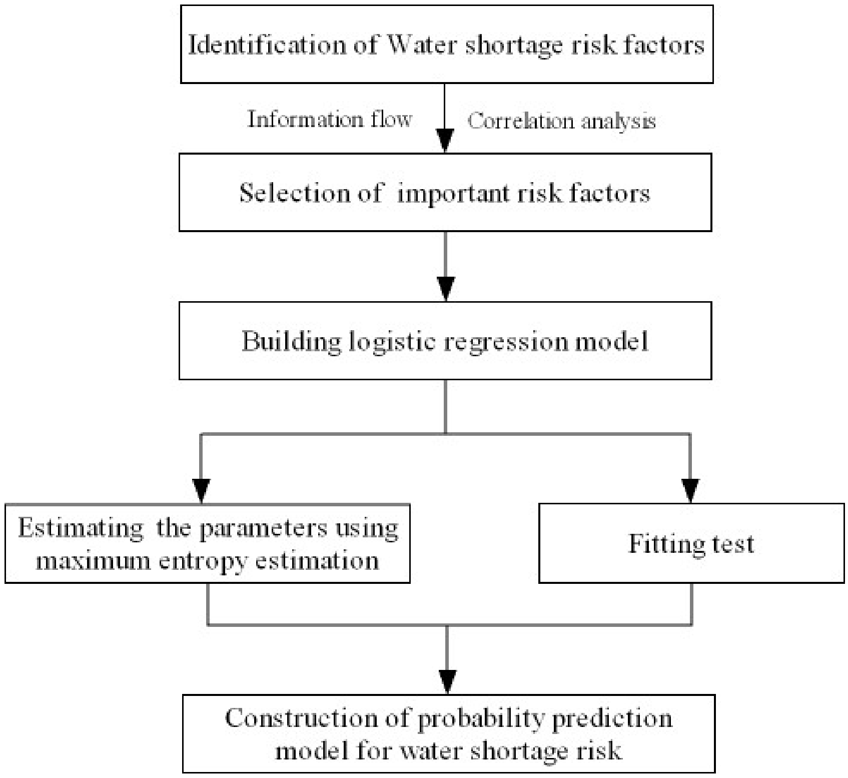

Figure 2. Flowchart showing the operation of the improved probability prediction model for water shortage risk

As can be seen from Figure 2 the model consists of a determination of water shortage risk factors and the construction of a logistic probability prediction model.

2.3.1 Identification of water shortage risk factors

Water shortage risk factors include characteristics of hazards and existing conditions of water resources vulnerability. Water resources vulnerability is referred to as the manifestation of the inherent states (e.g., physical, social, and ecological) of the water resources system that causes the system to be liable to a water shortage (Qian et al., 2016). According to the study of Plummer et al. (2012), there are 50 different water vulnerability assessment tools, and the water vulnerability indicators of these tools are quite different. Therefore, a universal standard understanding of water resource vulnerability indicators is difficult to develop. We established the indicators from perspective of hydrological conditions, water resources, water supply and water use. The risk factors are: precipitation (P), water resources per capita (W_p), water consumption per GDP (W_c), satisfactory rate of water demand (S_r), and utilization rate of water resources (U_r), proportion of industrial water use (IW_p), proportion of agricultural water use (AW_p), proportion of domestic water use (DW_p) and the treatment rate of domestic sewage (DS_r). These indicators are defined as follows (Qian et al., 2014):

$$W_p = \frac{W}{N} \quad (1)$$

where W is the total amount of water resources, and N is the population size.

$$W_c = \frac{the\ amount\ of\ water\ use}{GDP} \quad (2)$$

$$U_r = \frac{W_{ss} + W_{gs}}{W} = \frac{W_{as}}{W} \quad (3)$$

where W_{ss} is the surface water supply, W_{gs} is the groundwater supply, and W is the total amount of water resources.

$$DS_r = \frac{DS_t}{DS} \quad (4)$$

where DS_t is the amount of sewage treated and DS is the total amount of sewage discharged.

$$S_r = \frac{W_{as}}{W_{td}} \quad (5)$$

where W_{as} is the water supply, and W_{td} is the water demand.

$$IW_p = \frac{IW}{WU} \quad (6)$$

$$AW_p = \frac{AW}{WU} \quad (7)$$

$$DW_p = \frac{DW}{WU} \quad (8)$$

where IW is the industrial water use, AW is the agricultural water use, DW is the domestic water use and WU is total water use.

2.3.2 Selection of important risk factors

The purpose of this section was to select some important factors that have an significant impact on water shortage risk. Liang (2014) reported that the cause and effect between two time series can be measured by the time rate of information flowing from one series to the other. Liang proposed a concise formula for causal analysis. The causality is measured by information flow. Therefore, we can use the information inflow to unravel the cause-effect relation between the risk factors and water shortage risk.

According to Liang (2014), for series X_1 and X_2, the rate of information flowing (units: nats per unit time) from the latter to the former is

$$T_{2\to 1} = \frac{C_{11}C_{12}C_{2,d1} - C_{12}^{2}C_{1,d1}}{C_{11}^{2}C_{22} - C_{11}C_{12}^{2}} \quad (9)$$

where C_{ij} is the sample covariance between X_i and X_j, C_{i,d_j} is the covariance between X_i and $\dot{X}_j$,and $\dot{X}_j$ is the difference approximation of $\frac{dX_j}{dt}$ using the Euler forward scheme.

$$\dot{X}_{j,n} = \frac{X_{j,n+k} - X_{j,n}}{k\Delta t} \quad (10)$$

According to Liang (2014), with $k \geq 1$, for a general time series $k = 1$ would be suitable. If $T_{2\to 1} = 0$ or the absolute value of $T_{2\to 1}$ is less than 0.01, X_2 does not cause X_1, otherwise it is causal. A positive $T_{2\to 1}$ means that X_2 functions to make X_1

more uncertain, while a negative value means that X_2 tends to stabilize X_1. Liang (2015) proposed a method of normalizing the causality between time series and the range of value for $T_{2\to1}$ is 0 and 1.

2.3.3 Correlation analysis of selected risk factors

In theory, a probability prediction model requires variables to be mutually independent. Therefore, it is necessary to perform a correlation analysis. Because all of the factors are continuous variables, Pearson correlation coefficients are often applied. If the absolute correlation coefficient is greater than 0.5, there is a significant correlation between two factors.

2.4 Risk probability prediction model using maximum entropy estimation

A logistic regression model is a nonlinear regression method of studying a binary categorical or multi-categorical variable and its impact factors. Because a water shortage either occurs or does not occur, water shortage risk belongs to a binary categorical variable. Therefore, we can use a logistic regression model to simulate the relation between water shortage risk and its factors. Suppose the risk factors are $\left\{x_{ij}\left(i=1,2,\cdots,n;\, j=1,2,\cdots,m\right)\right\}$, where x_{ij} denotes the value of the *jth* factor in the *ith* year. The risk sequence is $\left\{y_i\left(i=1,2,\cdots,n\right)\right\}$, where $y_i=\begin{cases}0, & \textit{water shortage risk does not occur}\\ 1, & \textit{water shortagerisk occurs}\end{cases}$, and is the observed value of the ith year.

$p_i=p\left(y_i=1\middle|x_{ij}\left(j=1,2,\cdots,m\right)\right)$ is the conditional probability when y_i=1 under the conditions of $x_{ij}\left(i=1,2,\cdots,n;\ j=1,2,\cdots,m\right)$. The logistic regression model is

$$p_i=\frac{1}{1+e^{-(\alpha+\beta_1 x_{i1}+\beta_2 x_{i2}+\ldots+\beta_m x_{im})}} \tag{11}$$

where $\alpha,\beta_1,\beta_2,\cdots,\beta_m$ are the estimated parameters. The parameters are often determined by a maximum likelihood estimation. The log likelihood equation of computing $\alpha,\beta_1,\beta_2,\cdots,\beta_m$ is as follows:

$$\begin{cases}\dfrac{\partial L}{\partial \alpha}=\sum_{i=1}^{n}\left[y_i-\dfrac{\exp\left(\alpha+\sum_{j=1}^{m}\beta_j x_{ij}\right)}{1+\exp\left(\alpha+\sum_{j=1}^{m}\beta_j x_{ij}\right)}\right]=0\\ \dfrac{\partial L}{\partial \beta_j}=\sum_{i=1}^{n}\left[y_i-\dfrac{\exp\left(\alpha+\sum_{j=1}^{m}\beta_j x_{ij}\right)}{1+\exp\left(\alpha+\sum_{j=1}^{m}\beta_j x_{ij}\right)}x_{ij}\right]=0 \quad j=1,2,\cdots,m\end{cases} \tag{12}$$

According to Eq. (12), a large number of observed values of risk ($y_i\left(i=1,2,\cdots,n\right)$) and its factors are required for parameter estimation. Unfortunately, the correlated samples between risk and its controlling factors are insufficient. It is therefore far better to estimate the parameters. In this case, the maximum likelihood estimation is not applicable for parameter estimation. An alternative approach for parameter estimation is therefore required.

Thus, we proposed a new parameter estimation method based on the maximum entropy principle. The new method is named after maximum entropy estimation. The new method does not require the observed values of risk, and it requires only some observed values of the factors. Its principle is as follows.

For an observation, we can define its entropy to evaluate its degree of uncertainty.

According to Jones and Jones (2000), the entropy of the ith observation of water shortage risk is

$$
\begin{aligned}
H(p_i) &= -C\left[P_i \ln P_i + (1-P_i)\ln(1-P_i)\right] \\
&= -C\left[P_i \ln\left(\frac{P_i}{1-P_i}\right) + \ln(1-P_i)\right] \\
&= -C\left\{\frac{\left(\alpha + \sum_{j=1}^{m}\beta_j x_{ij}\right)}{1+\exp\left[-\left(\alpha + \sum_{j=1}^{m}\beta_j x_{ij}\right)\right]} - \ln\left(1+\exp\left(\alpha + \sum_{j=1}^{m}\beta_j x_{ij}\right)\right)\right\}
\end{aligned}
\tag{13}
$$

where C is a positive value and $p_i = p\left(y_i = 1 \middle| x_{ij}\,(j = 1, 2, \cdots, m)\right)$ is the conditional probability when y_i =1 under the conditions of $x_{ij}\,(i = 1, 2, \cdots, n;\ j = 1, 2, \cdots, m)$. According to the maximum entropy principle, if the values of $H(P_i)$ reaches a maximum, the optimal parameters are obtained (Jones and Jones, 2000). The reasons for obtaining a solution based on the maximum entropy principle are as follows. ① It conforms to the principle of entropy increase, which states that the entropy of an isolated system tends to reach a maximum. ② It accords with the principle that the solution should be in line with the sample/data and the least hypotheses must be constructed regarding the unknown parts when the data is insufficient. ③ It fits the maximum multiplicity principle. The multiplicity of a state refers to the number of possible ways in which a system can evolve to that state. The maximum multiplicity principle states that the greater the multiplicity of a state, the larger the possibility that a system is in this state.

2.4.1 Parameter estimation

Based on the analysis above, an optimization model can be constructed as follows:

$$\max\ H_i = -C\left\{ \frac{\left(\alpha+\sum_{j=1}^{m}\beta_j x_{ij}\right)}{1+\exp\left[-\left(\alpha+\sum_{j=1}^{m}\beta_j x_{ij}\right)\right]} - \ln\left(1+\exp\left(\alpha+\sum_{j=1}^{m}\beta_j x_{ij}\right)\right)\right\} \qquad (14)$$

According to the extreme theory of multivariate function (Khuri 2003), we can obtain

$$\begin{cases} \dfrac{\partial H_i}{\partial \alpha} = \dfrac{\left\{1+\exp\left[-\left(\alpha+\sum_{j=1}^{m}\beta_j x_{ij}\right)\right]\right\}+\left(\alpha+\sum_{j=1}^{m}\beta_j x_{ij}\right)\cdot\exp\left[-\left(\alpha+\sum_{j=1}^{m}\beta_j x_{ij}\right)\right]}{\left\{1+\exp\left[-\left(\alpha+\sum_{j=1}^{m}\beta_j x_{ij}\right)\right]\right\}^2} - \dfrac{\exp\left(\alpha+\sum_{j=1}^{m}\beta_j x_{ij}\right)}{1+\exp\left(\alpha+\sum_{j=1}^{m}\beta_j x_{ij}\right)} = 0 \\ \dfrac{\partial H_i}{\partial \beta_j} = \dfrac{x_{ij}\cdot\left\{1+\exp\left[-\left(\alpha+\sum_{j=1}^{m}\beta_j x_{ij}\right)\right]\right\}+x_{ij}\cdot\left(\alpha+\sum_{j=1}^{m}\beta_j x_{ij}\right)\cdot\exp\left[-\left(\alpha+\sum_{j=1}^{m}\beta_j x_{ij}\right)\right]}{\left\{1+\exp\left[-\left(\alpha+\sum_{j=1}^{m}\beta_j x_{ij}\right)\right]\right\}^2} - \dfrac{x_{ij}\cdot\exp\left(\alpha+\sum_{j=1}^{m}\beta_j x_{ij}\right)}{1+\exp\left(\alpha+\sum_{j=1}^{m}\beta_j x_{ij}\right)} = 0 \end{cases} \qquad (15)$$

The optimal estimation $\boldsymbol{\alpha}, \boldsymbol{\beta}_j\left(j=1,2,\cdots,m\right)$ can be obtained by solving Eq. (15). Numerical approaches are often used to obtain an approximate solution of Eq. (15) rather than its exact solution. Therefore, we made use of the optimization function of Matlab to estimate the parameters, i.e., the fminsearch function. If there are n observations, there are n $H_i\left(i=1,2,\cdots,n\right)$. It is impossible to find the parameters that make all the $H_i\left(i=1,2,\cdots,n\right)$ reach the maximum value. According to the maximum entropy principle, the greater the entropy is, the larger the uncertainty of an observation is. Therefore, the maximum value of the sequences $\left\{H_i, i=1,2,\cdots,n\right\}$ was taken as the objective function of the optimization model.

2.4.2 Goodness-of-fit test

According to Brown (1982), a goodness-of-fit test should be made for evaluating the fitting effect of the logistic regression model and its ability to identify water shortage risk. In this study, the Kolmogorov-Smirnov Test (K-S) test and Pearson χ^2 test are used.

2.4.2.1 K-S test (t)

A K-S test is often applied as a fitting test. It can be used to test the ability of the model to identify water shortage risk. The value of K-S is between 0 and 1; the greater the value is, the better the logistic model is. The idea is as follows.

Let $F_{n1}(x)$ be the cumulative probability distribution of the samples that do not encounter a water shortage. $F_{n2}(x)$ is the cumulative probability distribution of the samples that encounter a water shortage. A two independent samples test is then applied to compare whether the empirical distribution functions of two samples are the same. The test is as follows:

$$H_0: F_{n1}(x) = F_{n2}(x) \quad H_1: F_{n1}(x) \neq F_{n2}(x) \tag{16}$$

The value of K-S is:

$$K-S = \mathbf{max}\left|F_{n1}(x) - F_{n2}(x)\right| \tag{17}$$

When $N \to \infty$, the cumulative distribution curve and probability density curve of two samples can be obtained. The value of K-S is the maximum value of the cumulative distribution functions. When the value of K-S is greater than 0.35, the logistic regression model is applicable. The international classification standard of the logistic model is shown in Table 1 (Brown, 1982).

Table 1. The international classification standard of the logistic model

K-S	The effect of the model

<0.2	Bad
0.2~0.4	General
0.4~0.5	Good
0.5~0.6	Better
0.6~0.75	Very good
0.75~1	Perfect

2.4.2.2 Pearson χ^2 test

The test is as follows:

$$H_0: \text{the fitting is good} \qquad H_1: \text{the fitting is bad} \tag{18}$$

The expression of the χ^2 statistic is as follows.

$$\chi^2 = \sum_{j=1}^{l} \frac{\left(O_j - E_j\right)^2}{E_j} \tag{19}$$

where $j = 1, 2, \cdots, l$, l is the number of covariant types, O_j is the observed frequency of the *jth* covariant type, and E_j is the predicted frequency of the *jth* covariant type. The degree of freedom is the difference between the number of covariant types and parameters.

3 Results and discussion

In this section, a logistic probability prediction model for water shortage risk is constructed and discussed, and the risk probability in 2020 in Beijing is predicted using the proposed model.

3.1 Construction of the Logistic probability prediction model

A sequence of risk factors were obtained for the period from 1979 to 2012, and were computed based on Eqs. (1)~(8). The risk sequence $\left\{y_i\left(i=1,2,\cdots,34\right)\right\}$ from 1979 to 2012 was obtained as follows. According to Qian and Zhang et al. (2016), a water supply is deemed inadequate if the supply is less than the demand, leading to a water shortage in the water supply system. $y_i=\begin{cases}0, \textit{water shortage does not occur}\\1, \textit{water shortage occurs}\end{cases}$. Therefore, there are only 34-year data.

3.1.1 Determination of water resources vulnerability indicators

Based on the risk factors sequences from 1979 into 2012 (Table 2) and the method of normalized information inflow (Liang, 2015), the values of normalized information flow from the factors to risk are shown in Table 3. According to the normalized information flow results (Table 3), the value of the normalized information flow from AW_p to water shortage risk is only 0.0031, and it is very little. It was concluded that the AW_p does not result in a water shortage risk. Therefore, AW_p was removed as risk factors.

Table 2. The values of the risk factors and risk from 1979 to 2012

Year	W_c (m^3 per CNY)	W_p (m^3 per capita)	U_r	P (mm)	DS_r (%)	AW_p	DW_p	IW_p	S_r	Risk
1979	0.36	426.15	1.12	652.00	10.20	0.56	0.10	0.33	0.71	0
1980	0.36	287.52	1.94	387.30	9.40	0.63	0.10	0.27	0.41	1
1981	0.35	261.10	2.00	433.50	10.80	0.66	0.09	0.25	0.40	1
1982	0.30	391.44	1.29	585.10	10.90	0.61	0.10	0.29	0.62	1
1983	0.26	365.26	1.37	465.50	10.20	0.66	0.10	0.24	0.58	1
1984	0.18	407.36	1.02	442.10	10.00	0.55	0.10	0.36	0.79	0
1985	0.12	387.36	0.83	611.20	10.00	0.32	0.14	0.54	0.96	0
1986	0.13	262.94	1.35	560.30	8.90	0.53	0.20	0.27	0.59	1
1987	0.09	369.25	0.80	662.60	7.70	0.31	0.23	0.45	1.00	0
1988	0.10	369.27	1.08	594.70	7.40	0.52	0.15	0.33	0.74	0

1989	0.10	200.47	2.07	479.50	6.60	0.55	0.14	0.31	0.39	1
1990	0.08	330.20	1.15	662.40	7.30	0.53	0.17	0.30	0.70	0
1991	0.07	386.56	0.99	662.70	6.60	0.54	0.18	0.28	0.80	0
1992	0.07	203.63	2.07	500.00	1.20	0.43	0.24	0.33	0.39	1
1993	0.05	176.89	2.30	424.30	3.10	0.45	0.21	0.34	0.35	1
1994	0.04	403.73	1.01	727.70	9.60	0.46	0.23	0.32	0.79	0
1995	0.03	242.51	1.48	608.90	19.40	0.43	0.26	0.31	0.54	1
1996	0.02	364.22	0.87	669.40	21.20	0.47	0.23	0.29	0.92	0
1997	0.02	179.44	1.81	419.00	22.00	0.45	0.28	0.28	0.44	1
1998	0.02	302.67	1.07	687.40	22.50	0.43	0.30	0.27	0.75	0
1999	0.02	113.11	2.93	384.70	25.00	0.44	0.30	0.25	0.27	1
2000	0.01	123.64	2.40	446.60	39.40	0.41	0.33	0.26	0.33	1
2001	0.01	138.62	2.03	462.00	42.00	0.45	0.32	0.24	0.39	1
2002	0.01	113.13	2.15	413.00	45.00	0.45	0.34	0.22	0.37	1
2003	0.01	126.34	1.84	453.00	50.10	0.39	0.38	0.23	0.41	1
2004	0.01	143.36	1.52	539.00	53.90	0.39	0.39	0.22	0.50	1
2005	0.00	150.85	1.27	468.00	62.40	0.38	0.42	0.20	0.54	1
2006	0.00	154.97	1.14	448.00	73.80	0.37	0.45	0.18	0.57	1
2007	0.00	145.74	1.13	499.00	76.20	0.36	0.48	0.17	0.55	1
2008	0.00	201.77	0.74	638.00	78.90	0.34	0.51	0.15	0.78	0
2009	0.00	124.22	1.08	448.00	80.29	0.34	0.52	0.15	0.49	1
2010	0.00	117.64	0.99	524.00	81.00	0.32	0.42	0.14	0.52	1
2011	0.00	132.81	0.88	552.00	81.70	0.30	0.43	0.14	0.60	1
2012	0.00	190.89	0.58	708.00	83.00	0.26	0.45	0.14	0.88	0

According to Liang (2014), a positive value of the information flow means that the factor makes water shortage risk more uncertain, while a negative value means that the indicator tends to stabilize water shortage risk. Therefore, all the factors tend to make water shortage risk more uncertain. Furthermore, the impact of P, W_p, W_c are very significant.

Table 3. The values of information flow from the factors to water shortage risk

Factors	Information flow
W_c	0.3560

W_p	0.4823
U_r	0.3109
P	0.1575
DS_r	0.2413
IW_p	0.1320
AW_p	0.0031
S_r	0.1247
DW_p	0.1164

A correlation analysis was performed on the remaining factors. The values of the Pearson correlation coefficients are shown in Table 4.

Table 4. Pearson correlation coefficients for the relations between various factors

Pearson correlation coefficients	W_c	W_p	U_r	P	DS_r	DW_p	IW_p	S_r
W_c	1	0.603	0.047	–0.066	–0.559	0.354	–0.780	0.047
W_p	0.603	1	–0.455	0.571	–0.682	0.654	–0.753	0.696
U_r	0.047	–0.455	1	–0.723	–0.268	0.026	–0.157	–0.869
P	0.066	0.571	–0.723	1	–0.100	0.219	–0.064	–0.820
DS_r	–0.559	–0.682	–0.268	–0.100	1	–0.802	0.902	–0.087
DW_p	0.354	0.654	0.026	0.219	–0.802	1	–0.715	0.354
IW_p	–0.780	–0.753	–0.157	–0.064	0.920	–0.715	1	–0.013
S_r	0.047	0.696	–0.869	0.820	–0.087	0.354	–0.013	1

Based on the results in Tables 3 and 4, AW_p, S_r, IW_p, and DW_p were removed as risk factors. Therefore, the selected factors for logistic regression model were W_c, W_p, U_r, P and DS_r.

3.1.2 Construction of the logistic risk probability predication model

The data for the risk and selected factors (W_c, W_p, U_r, P and DS_r) from 1979 to 2012 (Table 2) are used to construct the logistic risk predication probability model. Because there is only 34 samples, it is impossible to estimate the parameters by the maximum likelihood estimation. Substituting the sequences of W_c, W_p, U_r P and DS_r from 1979 to 2012 (Table 2) into Eq. (14), the values of parameters obtained by maximum entropy estimation can be obtained. The estimated values for $\alpha, \beta_1, \beta_2, \cdots, \beta_5$ are 61.6386, 0.004, -0.1262, -12.4077, -0.012 and -29.0963.

Therefore, the logistic regression model based on the maximum entropy estimation is as follows:

$$\textit{Predicted probability} = \frac{1}{1+e^{-(61.6386+0.004W_c-0.1262W_p-12.4077U_r-0.012P-29.0963DS_r)}} \qquad (20)$$

Substituting the sequences of W_c, W_p, U_r, P and DS_r from 1979 to 2012 into Eq. (20), the predicted probability values of water shortage risk by the maximum entropy estimation is shown in Fig. 3.

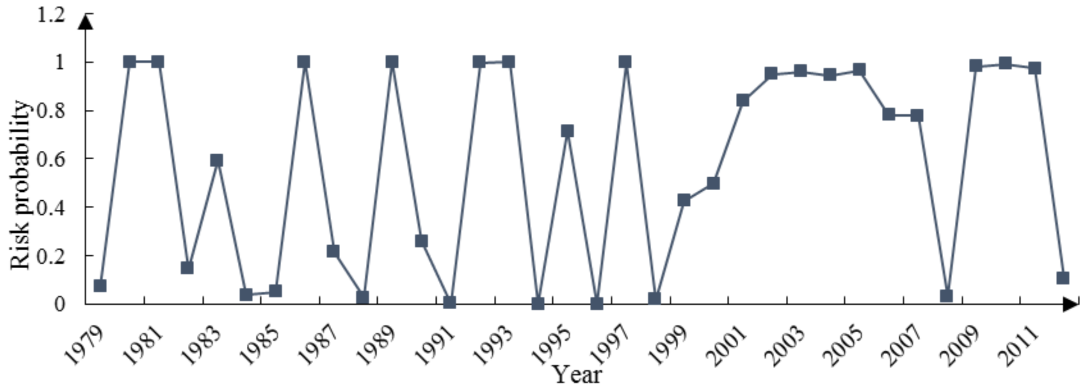

Figure 3. The predicted probability generated by the maximum entropy estimation from 1979 to 2012

If 0.5 is taken as threshold used to judge whether water shortage risk occurs, then the prediction accuracy by using the maximum entropy estimation can be obtained, and is shown in Tables 5. From Table 5, it can be seen that the average accuracy rate using the maximum entropy estimation was very high (91.18%). The maximum entropy estimation does not need observed values of risk ($y_i\left(i=1,2,\cdots,n\right)$), whereas the maximum likelihood estimation needs a large number of observed values of risk.

Table 5. The prediction accuracy using the maximum entropy estimation

	The prediction is that risk occurs	The prediction is that no risk occurs	Accuracy rate
Risk actually occurs	19	3	86.36%
Risk actually does not occur	0	12	100%
The average accuracy rate			91.18%

The K-S test and Pearson χ^2 test are performed and the results of the tests are obtained. The value of K-S is 0.955 and according to Table 1, the logistic probability prediction model was applicable. Moreover, the probability value was 0.000(i.e., less than 0.05), so the null hypothesis was rejected. Therefore, the ability of the logistic regression model to predict water shortage is very strong.

Substituting the observed frequency and the predicted frequency into Eq. (19), the value of the χ^2 statistics was 2.333 (the number of covariant type was 8). Because the number of parameters was 6, there were 2 degrees of freedom. The $\chi^2_{0.1}(2)$ was equal to 4.605 and was much greater than 2.333. Therefore, the null hypothesis was accepted, i.e., the fitting of the model was very good. Based on the results of the K-S test and Pearson χ^2 test, it was concluded that the model was applicable.

3.2 Risk probability prediction in 2020 in Beijing

3.2.1 Risk probability prediction (without considering the use of transferred and reclaimed water)

Because the inflow of 2020 is unknown, the inflow condition in 2020 was assumed to be any annual inflow conditions from 1956 to 2012. In this section we predict the risk probability of 2020 under different inflow conditions from 1956 to 2012. The sequences for risk factors (W_c, W_p, U_r, P and DS_r) were obtained and computed as follows. The precipitation in 2020 is assumed to be any annual precipitation from 1956 to 2012. First, an analysis of the balance between water supply and demand was performed and the sequences of water supply and demand under the inflow scenarios of 1956–2012 were obtained (Qian et al., 2016). The GDP of 2020 was the sum of the gross agricultural product, gross industrial product, and gross product of the third industry (details of the third industry are shown in Appendix A), using information taken from the literature, and was estimated to be 4711.852 billion CNY (Qian et al., 2016). N (the population size of 2020) was 24.43 million (Qian et al. 2016). The total amount of water resources from 1956 to 2020 were considered to consist of fifty-seven types of water resources in 2020. Substituting the total water resources sequences and N of 2020 into Eq. (1), the sequence of W_p could be computed. Substituting the water demand sequences and GDP of 2020 into Eq. (2), the sequence of W_c could be computed. Substituting the sequence of the total water resources and water supply for 2020 into Eq. (3), the sequence of U_r could be obtained. The DS_r of 2020 was about 90% (Beijing Municipal Development and Reform Commission and Beijing Municipal Bureau of Water Affairs, 2009).

Substituting the sequences of W_c , W_p , U_r , P and DS_r into Eq. (20), the probability that a water shortage risk will occur in 2020 under the inflow scenarios of 1956–2012 was predicted, and is shown in Figure 4.

In Figure 4, the horizontal axis represents the inflow conditions of 1956–2012. Figure 4 shows that in 2020, the water shortage risk probability exceeded 0.95 under 33 different inflow conditions (accounting for 63.5% of all the inflow conditions) and exceeded 0.5 under 38 different inflow conditions (accounting for 73.1% of all the inflow conditions). In summary, there was a high probability of a water shortage risk in 2020, although the probability was very low in some high precipitation periods.

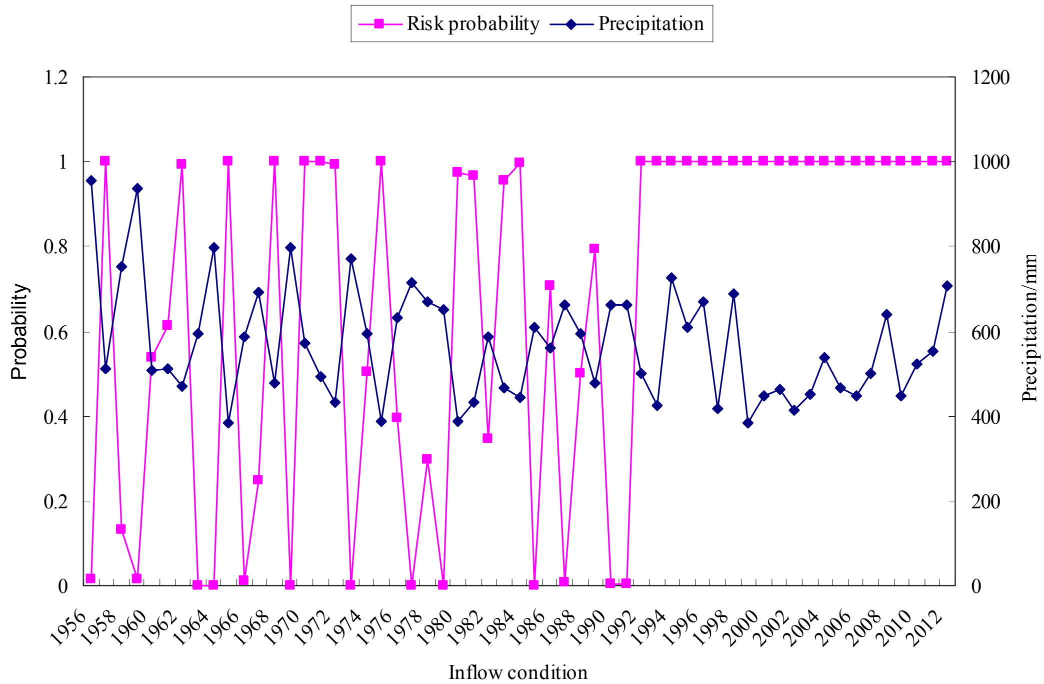

Figure 4. Risk probability under the inflow conditions of 1956–2012

3.2.2 Risk probability prediction after using transferred and reclaimed water

According to Qian et al. (2016), 1.05 billion m^3 of water will have been transferred to Beijing in 2020 and the amount of reclaimed water used may reach 1 billion m^3. After using transferred and reclaimed water, the total amount of water resources would increase, W_p and U_r would change and other indicators would remain unchanged. Therefore, the sequences of W_p and U_r under the inflow scenarios of 1956–2012 had to be computed again. Substituting the sequences of W_c, W_p, U_r, P and DS_r into Eq. (20), the water shortage risk probability in 2020 under the inflow scenarios of 1956–2012 (after using transferred and reclaimed water) was predicted, and the results are shown in Figure 5.

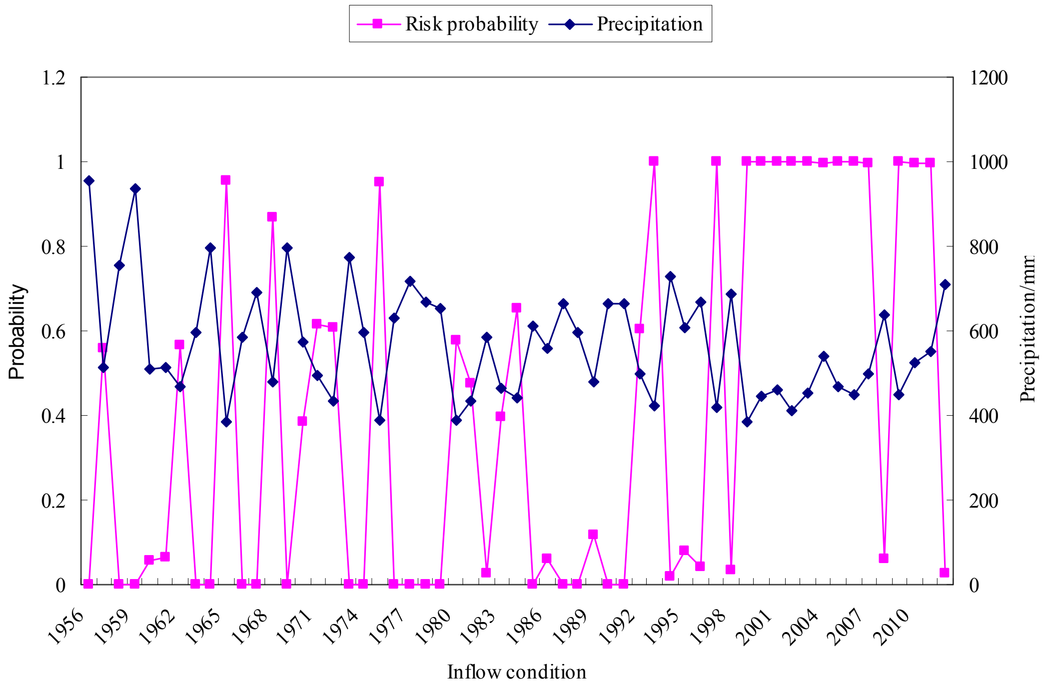

Figure 5. Values of risk probability under the inflow conditions of 1956-2012 after using transferred and reclaimed water

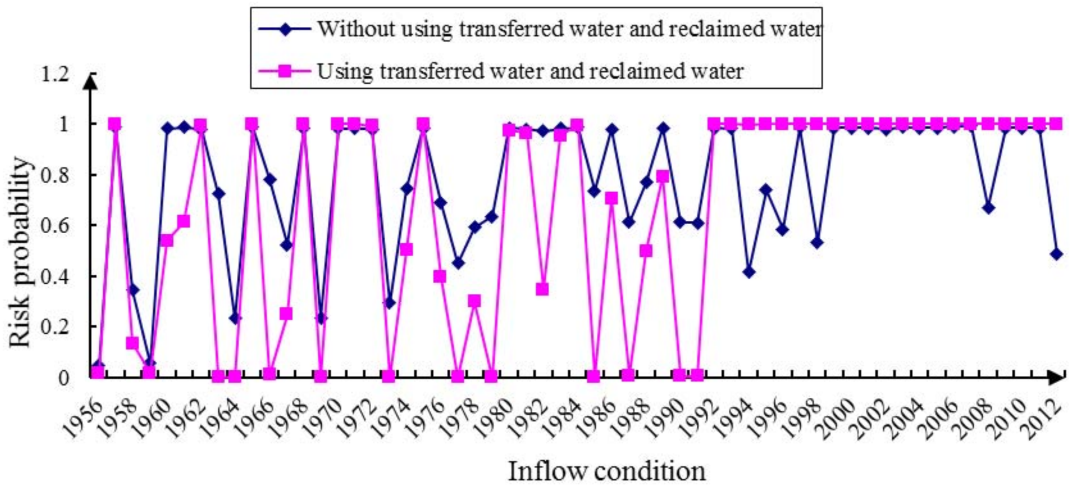

Figure 6. Comparison of risk probability before and after using transferred and reclaimed water

From Figures 5 and 6, it was concluded that the water shortage risk probability would decline under all inflow conditions (59.1% on average). However, the water shortage risk probability would still be high in some low inflow conditions. The risk probability exceeded 0.5 under 24 different inflow conditions (accounting for 46.2% of all inflow conditions). For example, the water shortage risk probability reached 1 under the inflow conditions of 1999–2008.

According to Qian et al. (2016), since 1999, Beijing has experienced drought in ten consecutive years. This has had a strong effect on the water resources of Beijing, including a significant reduction in surface water and severe over-exploitation of groundwater. This means that a water shortage may occur in 2020 under the inflow conditions of 1999–2008 although some measures have been taken. Moreover, water resources vulnerability was still high in 2020 after using transferred and reclaimed water (Qian et al., 2016). Therefore, we concluded that the water shortage risk probability would still be high in 2020 after using transferred and reclaimed water, especially in the case of precipitation deficits.

4 Conclusions

This study developed an improved logistic probability prediction model for water shortage risk in situations when there is insufficient data. The model consists of the following steps:

(1) Information flow was used to select some important factors that were likely to have a significant impact on water shortage risk. This could determine the cause-effect relation between the water shortage risk and its factors.

(2) The logistic regression model was applied to describe the nonlinear relation between water shortage risk and its factors. A new parameter estimation method based

on the entropy principle, i.e. maximum entropy estimation, was proposed for parameter estimation when insufficient data is available.

The results of the study were as follows. In 2020, the probability that a water shortage risk will occur exceeded 0.95 under 33 different inflow conditions (accounting for 63.5% of all inflow conditions) and exceeded 0.5 under 38 different inflow conditions (accounting for 73.1% of all inflow conditions). After using the transferred and reclaimed water, the water shortage risk probability declined under all inflow conditions (by 59.1% on average), but the water shortage risk probability was still high for some low inflow conditions. Risk probability exceeded 0.5 under 24 different inflow conditions (accounting for 46.2% of all inflow conditions).

However, some problems still exist with regard to the maximum entropy estimation. Initial values of the parameters should be given for the optimization function, but the optimization function belongs to local optimization, which was very sensitive to the initial values. Therefore, we may obtain an unsatisfactory result if the initial values are not correct. How best to search for a global optimum is an important and difficult issue, and will be the focus of our further study.

Appendix A. Glossary used in this paper

1. **Logistic regression model.** It is nonlinear regression method of studying binary categorical or multi-categorical variable and its impact factors.

2. **Maximum likelihood estimation**. It is a method of parameter estimation in statistics.

3. **Maximum entropy estimation**. We propose a new parameter estimation method for a logistic regression model when insufficient data is available. We called this new method maximum entropy estimation.

4. **Backward**. It is a method of selecting the variables for a logistic regression model. The methods of selecting the variables for a logistic regression model include enter, forward and backward.

5. **Information flow**. Information flow, proposed and named by Liang (2014), is a method for unraveling the cause-effect relation between time series.

6. **The extreme theory of multivariate function**. This is a theory used for calculating extreme values in advanced mathematics.

7. **Two independent samples test**. This is one type of Kolmogorov-Smirnov (K-S) test. The K-S test includes a one-sample K-S test, two independent sample test, and a test for several independent samples.

8. **The third industry**. In China, the third industry is also known as the service industry, and includes the traffic and transportation industry, communication industry, and commercial industry.

Appendix B. Abbreviations used in this paper

1. **PLA** People's Liberation Army of China.

2. **GDP** Gross domestic product

3. P. Precipitation.

4. W_p. Water resources per capita

5. W_c Water consumption per 10 thousand CNY GDP

6. S_r Satisfactory rate of water demand

7. U_r Utilization rate of water resources

8. IW_p Proportion of industrial water use

9. AW_p Proportion of agriculture water use

10. DW_p Proportion of domestic water use

11. DS_r Treatment rate of domestic sewage

12. **CNY**. The Chinese Yuan

13. **K-S test**. Kolmogorov-Smirnov Test

Acknowledgments The study was supported by National Natural Science Foundation of China (Grant Nos. 51609254, 51279006 and 51479003).

References

Balakrisheaa, N.: Approximate MLES for the location scale parameters of the half-Logistic distribution with type right-censoring, IEEE Transactions on Reliability, (40), 140–145, 1991.

Bedford, T., and Cooke, R.M.: Probability Risk Analysis: Foudations and Methods, Cambridge University Press, Cambridge, 2001.

Beijing Municipal Development and Reform Commission and Beijing Municipal Bureau of Water Affairs: Beijing City comprehensive planning of water resources, China Water Power Press, Beijing, 2009. (in Chinese)

Breslow, N.E. and Zaho, L.P.: Logistic regression for stratified case-control studies, Biometrics (44):891–899, 1988.

Brown, C.C.: On a goodness-of-fit test for the logistic model based on score statistics, Communications in Statistics, 11(10), 1087-1105, 1982.

Christodoulou, S.E.: Water resources conservancy and risk reduction under climatic instability, Water Resources Management 25, 1059–1062, 2011.

Giacomelli, P., Rossetti, A., and Brambilla, M.: Adapting water allocation management to drought scenarios, Nat. Hazards Earth Syst. Sci., 8, 293–302, 2008.

Giannikopoulou, A.S., Gad, F.K., and Kampragou, E.K.: Risk-based assessment of drought mitigation options: the case of Syros Island, Greece, Water Resources Management, 31(2), 655–669, 2017.

Haimes, Y.Y.: On the definition of vulnerability in measuring risks to infrastructures, Risk Analysis, 26(2), 293–296, 2006.

Haimes, Y.Y.: On the complex definition of risk: a systems-based approach, Risk Analysis, 29(12), 1647–1654, 2009

Hashimoto, T., Stedinger, J.R., and Loucks, D.P.: Reliability, resiliency and vulnerability criteria for water resources system performance evaluation, Water Resources Research, 18(1), 14–20, 1982.

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

Karimi, I., and Hüllermeier, E.: Risk assessment system of natural hazards: a new approach based on fuzzy probability, Fuzzy Sets and Systems, 158, 987–999, 2007.

Khuri, A.I.: Advanced calculus with applications in statistics, John Wiley & Sons, Inc., Hoboken, New Jersey, 2003.

Li, F.W., Qiao, J.L., Zhao, Y., and Zhang, W.: Risk assessment of groundwater and its application part II: using a groundwater risk maps to determine control levels of the groundwater, Water Resources Management, 28(13), 4875–4893, 2014

Liang, X.S.: Unraveling the cause-effect relation between time series, Physical Review E, 90, 052150-1–052150-11, 2014.

Liang, X.S.: Normalizing the causality between time series, Physical Review E, 92, 022126, 2015

Mackenzie, A.C.: Summarizing risk using risk measures and risk indices, Risk Analysis, 34(12), 2143–2162, 2014.

Plummer, R., Loë de Rob, and Armitage, D.: A systematic review of water vulnerability assessment tools, Water Resources Management, 26, 4327–4346, 2012.

Qian, L.X., Wang, H.R., and Zhang, K.N.: Evaluation criteria and model for risk between water supply and water Demand and its application in Beijing, Water Resources Management, 28, 4433–4447, 2014.

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

Wang, C.H., and Blackmore, J.M.: (2012) Supply–demand risk and resilience assessment for household rainwater harvesting in Melbourne, Australia, Water Resources Management 26(15), 4381–4396, 2012.

Rajagopalan, B., Nowak, K., Prairie, J., Hoerling, M., Harding, B., Barsugli, J., Ray, A., and Udall, B.: Water supply risk on the Colorado River: Can management mitigate? Water resources research, 45, W08201, 2009.

Sandoval-Solis, S., McKinney, D.C., and Loucks, D.P.: Sustainability index for water resources planning and management, Journal of Water Resources Planning and Management, 137(5), 381–390, 2011.

Statistical Bureau of Beijing City: Statistical Yearbook 2014 of Beijing City. China Statistics Press, Beijing, 2014. (in Chinese)

Tidwell, V.C., Cooper, J.A., and Silva, C.J.: Threat assessment of water supply systems using markov latent effects modeling, Journal of Water Resources Planning and Management, 131(3), 218–227, 2005.

UNISDR: Terminology on disaster risk reduction, United Nations, Geneva, 2009.

Villagrán, De León J: Vulnerability: A conceptual and methodological review, SOURCE No.4.UNU-EHS, Bonn, 2006.

Weng, B.S., Yan, D.H., Wang, H., Liu, J.H., Yang, Z.Y., Qin, T.L., and Yin, J.: Drought assessment in the Dongliao River basin: traditional approaches vs. generalized drought assessment index based on water resources systems. Nat. Hazards Earth Syst. Sci., 15, 1889–1906, 2015.

Yang, C.C., Yeh, C.H., and Ho, C.C.: Systematic quantitative risk analysis of water shortage mitigation projects considering climate change, Water Resources Management, 29(4), 1067–1081, 2015.