Analysis and simulation of the influencing factors on regional water use based on information entropy

Lihua Ma\textsuperscript{a,b}, Shaozhong Kang\textsuperscript{c}, Yunqing Xuan\textsuperscript{b}, Xiaoling Su\textsuperscript{a} and Dimitri P. Solomatine\textsuperscript{d,e}

\textsuperscript{a}Northwest A\&F University, Yangling, Shaanxi, China
\textsuperscript{b}College of Engineering, Swansea University, Swansea, UK
\textsuperscript{c}Corresponding author. Center for Agricultural Water Research in China, China Agricultural University, Beijing 100083, China. E-mail: kangshaozhong@tom.com
\textsuperscript{d}UNESCO-IHE Institute for Water Education, Delft, The Netherlands
\textsuperscript{e}Water Resources Section, Delft University of Technology, Delft, The Netherlands

Abstract

With regional socio-economic development, the gap between water use demand and available water resources in arid and humid semi-arid areas becomes increasingly serious. In this study, the size of regional water use in the Guanzhong region and Shiyang river basin in northwest China are analyzed to identify important factors affecting it, with the aim of providing better and optimal development planning for the region. Information entropy is used to measure and characterize the diversity of regional water use. Agricultural development and meteorological factors are found to be the main issues affecting regional water use in both regions. A multiple-linear regression (MLR) model was built by applying correlation coefficient (R) and mutual information (MI) scores in the process. Results show that the low value of information entropy of water use in the Shiyang river basin is due to the high proportion of agriculture water use. Using input factors chosen by MI score was found to be the best model to simulate the change of regional water use in both regions. A method using an MLR model together with MI is shown to be able to quantify the relationship between the influencing factors and water use diversity with limited available data.

Keywords: Information entropy; Mutual information; Water use proportion

1. Introduction

The water use proportion is that proportion of the total volume of groundwater and surface water withdrawn for human use, and it is directly determined by the demand for water use for human


© IWA Publishing 2012
production and life (Lin & Wang, 2003). Understandably, an increasing demand for human production and life has always concerned water resource managers.

Various classifications of water use have been proposed by researchers and institutions, focusing on different aspects. For instance, the State Hydrological Institute (1998) indicated that water use could be classified into municipal, industrial, reservoir and agricultural water use. In China, water use is often classified as municipal, industrial and agricultural water use (Lin & Wang, 2003), with the proportion expressed as a percentage of water use for each sector. Although statistics of this kind can provide useful explanations of water use for particular regions, it is not easy to make comparisons between regions, as both the composition of water use and socio-economic development can vary dramatically, which necessitates the use of more sophisticated methods instead of a simple classification.

Setting each water use sector as one random variable and then combining this information together will help identify the trend of regional water use proportion changes. If the names of various water use components are disregarded, the problem of seeking a unified measurement in the trends of regional water use proportions is similar to that arising in determining the information content of the system, such as measuring the rate of creation of information in random phenomena, with the help of the concept of information entropy (Shannon, 1951). This concept has been widely used in many other areas, for instance, to indicate the complexity of the Mississippi river system (Li & Zhang, 2008) and to determine the evolution of water quality (Zou et al., 2006; Liu et al., 2010). It also serves our purpose well in terms of providing a ‘standard’ measurement of water use diversity and evolution trends. Furthermore, it also helps our next effort, i.e., uncovering the influencing factors that cause these differences, which are revealed by the information entropy measure.

In addition to investing efforts in quantifying the difference in water use patterns in different regions using method such as information entropy, it is also interesting (and necessary) to identify the so-called influencing factors that cause the differences. This is of special importance when helping local authorities improve their future planning by addressing water scarcity problems with clear targets, e.g., by mitigating the impact caused by the most influential factors. The mechanisms of how these factors impact on regional water use, however, are rather complex since most of the factors (such as climate change, land use change, water cycle process, and water conservancy engineering, for example) can also interact with one another (Conway et al., 1996; Vörösmarty et al., 2000). Changes in water use also depend on the characteristics of the different water sectors. In principle, municipal water use tends to increase with increasing population (Alcamo et al., 2003); agricultural water use is related to crop water requirements, crop area and rural population (Allen et al., 1998); and industrial water use usually increases with the increase of gross domestic product (Oki & Kanae, 2006). The impacts of these factors on regional water use are difficult to quantify and separate analytically, and therefore a statistical method able to reveal the relationship between the water use (the target) and the influencing factors can better represent the mechanism. However, we still need to differentiate the most important influencing factors from the rest, and this will be of great value in the decision-making process (allowing better planning, for example).

For the simulation, the multiple-linear regression (MLR) method can produce a linear equation that expresses the dependent variable as a function of several independent variables. The method has been widely used by many researchers, as it can be easily implemented yet is able to give a meaningful explanation of the processes involved. The inputs selection in building a MLR model is a key procedure and stepwise regression is one of the popular methods that can be used without empirical knowledge. In addition, correlation (Lachniet & Patterson, 2006) and principal component scores (Camdevyren
et al., 2005) are also used to choose inputs variables. Yokoo et al. (2001) pointed out that peak runoff was simulated well when the correlation coefficient has a large value. May et al. (2008) introduced the linear correlation coefficient (R) and mutual information (MI) methods for the effective selection of a model input. The MI method can be further helpful in determining best model inputs (Kwak & Choi, 2002; Peng et al., 2005). In this study, a similar approach was used to build the models for two typical areas in northwest China.

As pointed out by Kang et al. (2004), having limited water resources is one of the major challenges to the communities in the arid and semi-arid regions of northwest China. A better understanding of the influencing factors on regional water use is believed to be one of the key steps towards addressing the dilemma of socio-economic development and limited water resources. In this study, two typical areas, the Guanzhong area and the Shiyang river basin, which have different climates and socio-economic backgrounds, were chosen to build regional water use models that are able to address regional differences due to water use structures as well as the different influencing factors. The study involves several steps: (1) to analyze the change of regional water use by calculating information entropy; (2) to select the important environmental variables using linear correlation coefficients and mutual information methods; and (3) to build a MLR model with different selection variables (input) and regional information entropy of water use (output) to obtain the best model simulating the change of regional water use.

2. Data and methods

2.1. Study area

The Shiyang river basin (N 36°29′–39°27′; E 101°41′–104°16′) is located in the inland arid region of the Gansu province of northwest China (Figure 1(a)). The basin has a drainage area of 41,600 km² and lies in a typical continental temperate climate zone with low annual precipitation (less than 150 mm), high evaporation and strong solar radiation. The annual temperature of the Shiyang basin is about 8.8 °C. In 2008, it had 2.26 million inhabitants and a gross domestic product (GDP) of 387.32 × 10⁸ RMB.

![Fig. 1. Location of the Guanzhong region and the Shiyang river basin.](https://iwaponline.com/wp/article-pdf/14/6/1033/406142/1033.pdf)
The Guanzhong region (N 33°35′–35°50′; E 106°18′–110°37′) lies in the inland humid semi-arid region, in the Shaanxi province of northwest China (Figure 1(b)). The region has an area of 55,400 km² and it has a warm temperate continental monsoon climate with an annual temperature of 13.5 °C and annual precipitation of 670 mm. In 2008, the population of the region was 23.39 million and regional GDP was 4344.66 × 10^8 RMB.

2.2. Data collection

According to convention in China, water use is classified into municipal, industrial and agricultural water use (Lin & Wang, 2003). Data on water use were extracted from the Gansu and Shaanxi Provincial Water Resources bulletins (Shiyang river basin: 1994–2008; Guanzhong: 1987–2008). The variables included are population (10^4 person), urbanization (urban population proportion percentage), grain output (10^4 ton), GDP (10^8 RMB Yuan), agricultural output (10^8 RMB Yuan), industrial output (10^8 RMB Yuan), effective irrigation area (10^4 ha) and cultivated land (10^4 ha), all being socio-economic variables, and precipitation (mm), temperature (°C) and reference evapotranspiration (ET0, mm) as meteorological variables. The economic data were obtained from the Annual Statistics Reports from the Shaanxi and Gansu Provinces (Shiyang river basin: 1959–2005; Guanzhong: 1986–2005). The meteorological observation data were collected from the national weather station (Shiyang river basin: 1959–2005; Guanzhong: 1986–2005); while ET0 was calculated according to the Penman–Monteith equation by Tong et al. (2007) and Cao et al. (2007).

2.3. Methodology

2.3.1. Information entropy of water use. The information entropy of water use represents the proportions of agricultural, industrial and municipal water use, as a total measure. The proportions of agricultural (P1), industrial (P2) and municipal water use (P3) were taken as the random variables. When \( \sum_{i=1}^{n} P_i = 1 \), the information entropy of water use (Y) is (Singh, 1997):

\[
Y = -\sum_{i=1}^{n} P_i \log_2 P_i
\]  

(1)

The most important properties of the entropy measure are: (1) any tendency toward equalization of \( P_i \) for a given \( n \) increases \( Y \); (2) when \( P_i \) are all equal, the greater the value of \( n \), the greater the value of \( Y \); (3) \( 0 \leq Y \leq \log_2 n \); (4) the measure is defined for conditional probabilities, including posterior probabilities determined by Bayes’ theorem; and (5) entropy can be separated into ‘between’ and ‘within’ groups entropies.

In this study, \( n = 3 \). According to Equation (1), the information entropy of water use reaches its maximum value when \( P_1 = P_2 = P_3 \), i.e. \( Y_{\text{max}} = \log_2(n) = \log_2 3 \). It can also be seen that the information entropy calculated for ‘balanced’ water use, in which all sectors have similar proportions, is considerably higher than that for ‘unbalanced’ water use which is dominated by a couple of sectors with larger proportions.
2.3.2. Analysis methods of influencing factors. The mutual information and correlation coefficient were calculated as the measuring scores used for the analysis of the variables influencing the change of the information entropy of regional water use, from which a higher value of the score indicates a stronger dependence. There were three steps: (1) the potential influencing factors were selected according to the empirical method (see section 2.2); (2) the mutual information and correlation coefficient for the potential influencing variables and the water use were calculated, and then the ranking of all affecting variables was undertaken for the two methods; and (3) the results in step 2 were compared to obtain the most important variables influencing the difference of water use between the two regions after taking the local conditions into consideration.

2.3.2.1. Mutual information (MI). Assuming that the time series of potential influencing variables is \( X \) and the time series of the information entropy of water use is \( Y \), MI can be calculated using the following equation (Fraser & Swinney, 1986):

\[
\text{MI} = \int f_{X,Y}(x, y) \ln \left( \frac{f_{X,Y}(x, y)}{f_X(x)f_Y(y)} \right) \, dx \, dy
\]

(2)

where \( f_X(x) \) and \( f_Y(y) \) are the probability distribution functions of variable \( X \) and \( Y \) respectively; \( f_{X,Y}(x, y) \) is a joint (bivariate) marginal distribution function of \( X \) and \( Y \). When \( X \) and \( Y \) are independent, the joint probability density \( f_{X,Y}(x, y) \) would equal \( f_X(x)f_Y(y) \). The MI score in Equation (2) would, in that case, equal 0 (the ratio of the joint and marginal densities being one, giving the logarithm a value of zero). A high value of MI score would indicate a strong dependence between the two variables. Given any bivariate samples, Equation (2) can be converted into Equation (3) to calculate MI in a discrete fashion (Sharma, 2000).

\[
\text{MI} = \frac{1}{N} \sum_{i=1}^{N} \ln \left( \frac{P_{X,Y}(x_i, y_i)}{P_X(x_i)P_Y(y_i)} \right)
\]

(3)

where \( x_i \) and \( y_i \) are the \( i \)th bivariate sample pair in a sample of \( N \); and \( P_X(x_i) \), \( P_Y(y_i) \) and \( P_{X,Y}(x_i, y_i) \) are the respective univariate and joint probability densities estimated at the sample data points.

The key to successful computation of MI in Equation (3) is the accurate estimation of these probability distributions from a finite set of examples (Da Costa Couto, 2009). The original MI algorithm approximately calculates the probability densities using rough measures, such as a histogram with fixed bin width (Fraser & Swinney, 1986; Zheng & Billings, 1996). More recently, kernel density estimation techniques have been used because of their stability, efficiency and robustness. For example, the Gaussian kernel function (Scott, 1992) can be used, as given by:

\[
\hat{f}_X(X) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{(2\pi)^{d/2} \lambda^d \det(S)^{1/2}} \times \exp \left( -\frac{(x - x_i)^T S^{-1}(x - x_i)}{2\lambda^2} \right)
\]

(4)

where \( \hat{f}_X(X) \) is the multivariate kernel density estimate of the \( d \)-dimensional variable set \( X \) at coordinate location \( x \); \( x_i \) is the \( i \)th multivariate data point, for a sample of \( N \); \( S \) is the sample covariance matrix of the
variable set $X$; $\det(S)$ represents the determinant operation; and $\lambda$ is the smoothing parameter, known as the bandwidth of the kernel density estimate.

Smaller values of $\lambda$ tend to lead to density estimates that give too much emphasis on individual points. Larger values of $\lambda$ tend to over-smooth the probability density with all detail, spurious or otherwise, becoming obscured. Various rules-of-thumb are available in the literature to help choose an optimal value for the bandwidth ($\lambda$). Scott (1992) used the Gaussian reference bandwidth because it was relatively simple and computationally efficient, as follows:

$$\lambda_{\text{ref}} = \left( \frac{4}{d+2} \right)^{1/(d+4)} N^{-1/(d+4)}$$

(5)

where $N$ and $d$ are the numbers and dimension of the sample in the multivariate variable set, respectively.

2.3.2.2. Correlation coefficient. The correlation coefficient ($R$) is commonly used to quantify the strength of the linear association between two independently measured variables ($X$ and $Y$). In this study, $X$ means potential influencing variables; $\bar{X}$ represents the mean of the variables; the information entropy of water use is $Y$; $i$ represents the time series from 1994 to 2005, $n$ equals 12. $R$ can then be defined as:

$$R = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2(Y_i - \bar{Y})^2}}$$

(6)

The 5% significance level of the correlation coefficient with 12 samples is $R_{0.05}$ ($R_{0.05} = 0.5760$). Thus all 11 variables above would be judged significant. The higher the value of $R$, the stronger the relationship between $X$ and $Y$.

Note that while the mutual information between the two variables strongly suggests that the variables are dependent, the reverse is true when the correlation coefficient is used. It is obvious that the non-linearity in the sample makes it impossible for a method such as correlation to recognize the nature of the dependence. The density-based formulation of the MI criterion enables accurate characterization of the nonlinear dependence pattern.

2.3.3. Simulation models. Three MLR models were built to quantify the relationship between the selected variables and the information entropy of water use. The models were used to: (1) compare different methods of finding influencing variables; (2) compare the change of information entropy of water use in the Shiyang river basin and the Guanzhong region in recent years.

Depending on how the input variables were chosen, the three MLR models were named the $R$-model (which takes four variables with highest correlation coefficients ($R$)), the MI-model (which chooses four variables with highest MI values) and the Full_model (which utilizes all 11 variables as the inputs). The simulated value of information entropy of water use, $\hat{Y}_i$, for the $i$th year can be calculated for these three models using Equations (7)–(9), respectively:

$$\hat{Y}_i = \beta_{0,R} + \sum_{j=1}^{4} \beta_{j,R}X_{j,R}$$

(7)
\[ Y_i = \beta_{0_{MI}} + \sum_{j=1}^{4} \beta_{j_{MI}} X_{j_{MI}} \]  
(8)

\[ \hat{Y}_i = \beta_0 + \sum_{j=1}^{11} \beta_j X_j \]  
(9)

where \( \beta_{R_j} \) is the coefficient in the R_model, and \( X_{j_R} \) means the \( j \)th selected factor by the correlation coefficient method, when \( j = 1, 2, 3, 4 \); \( \beta_{MI} \) is the coefficient in the MI_model, and \( X_{j_{MI}} \) means the \( j \)th selected factors by mutual information method, when \( j = 1, 2, 3, 4 \); \( \beta \) is the coefficient in the Full_model, and \( X_j \) means the \( j \)th factors, when \( j = 1, 2, \ldots 11 \).

The dataset used in this study only has 12 samples (1994–2005), thus additional methods are applied to ensure model performance. The \( K \)-fold cross validation method was used to estimate the generalization error (Kohavi, 1995; Boyce et al., 2002; Huang et al., 2007); in \( K \)-fold cross validation, the training data (which are the data used to determine the model coefficients) are randomly split into \( K \) mutually exclusive subsets (folds) of approximately equal size. Leave-one-out cross validation (LOO CV) can be viewed as an extreme form of the \( K \)-fold cross validation in which \( K \) is equal to the number of examples. It is known that the LOO procedure gives an almost unbiased estimate of the expected generalization error (Luntz & Brailovsky, 1969). In this study, the training data were split into three (\( K = 3 \)) and 12 (LOO) mutually exclusive subsets (folds) of approximately equal size.

3. Results and discussion

3.1. Variation of the information entropy of water use

The information entropy of water use in the Shiyang river basin and Guanzhong region is shown in Figure 2. During the period 1994–2008, the information entropy of water use in the Guanzhong region (0.798–0.941) was much higher than for the Shiyang river basin (0.249–0.471), indicating that there was significant difference in water use between the two regions.

It is worth noting that the information entropy of water use, according to its formulation (Equation (1)), is able to reflect the diversity of water use across different sectors, in our case, \( P_1 \), \( P_2 \) and \( P_3 \) (agricultural use, industrial use and municipal use); however, the individual sector contribution cannot be distinguished by the information entropy itself. Nevertheless, information entropy is a very useful tool to analyze and compare water use from various regions using information (such as the percentage of sectoral water usage) that can easily be obtained.

As a reference, worldwide information entropy of water use is shown in Table 1. The original data come from UNESCO Statistics (2001; see State Hydrological Institute (1998)), and the results were calculated according to Equation (1). The results of information entropy in Europe, North America, and in Australia and Oceania did not change significantly during the period 1900–2000, and the value remained high over the years (\( Y_{\text{min}} = 0.865; \ Y_{\text{max}} = 1.098 \)). By contrast, information entropy in Africa and Asia stayed low (\( Y_{\text{min}} = 0.119; \ Y_{\text{max}} = 0.594 \)) although there was a significant increase during the period 1900–2000. The information entropy of South America grew gradually from 0.424 to 0.930.
For Asian and African countries in the 1900s, water use was dominated by agriculture, and thus the higher proportion of agricultural water use and the lower proportion of industrial and municipal water use produced a lower information entropy of water use. With the social and economic development in these areas, industrial water use increased (with industrial development) and municipal water use also increased (alongside developing urbanization). As a result, information entropy increased gradually. In contrast, since Europe, North America, and Australia and Oceania had almost completed their transition from an agricultural society to an industrial society in 1900 (due to the Industrial Revolution), the information entropy has remained high ever since. As for the period 1900–2000, due to the joint impact of more effective water use in industry as well as water reuse technologies, a very mild increase or even slight decrease is observed in the proportion of industrial and municipal water use, e.g., as seen in the Kuznets growth curve in industrial water use (Jia et al., 2006). But still the proportions of the different sectors were maintained close to one another, and the information entropy of water use as such was able to keep higher value over the years.

Table 1. Comparison of information entropy of water use between the study region in northwest China and other regions in the world. (Original data taken from State Hydrological Institute (1998); results were then calculated according to Equation (1)).

<table>
<thead>
<tr>
<th>Region</th>
<th>Year</th>
<th>Europe</th>
<th>North America</th>
<th>Africa</th>
<th>Asia</th>
<th>South America</th>
<th>Australia &amp; Oceania</th>
<th>World</th>
<th>Guanzhong region</th>
<th>Shiyang river basin</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1900</td>
<td>1.098</td>
<td>0.865</td>
<td>0.160</td>
<td>0.119</td>
<td>0.461</td>
<td>0.865</td>
<td>0.424</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1950</td>
<td>1.080</td>
<td>0.907</td>
<td>0.372</td>
<td>0.288</td>
<td>0.584</td>
<td>0.900</td>
<td>0.646</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1980</td>
<td>0.983</td>
<td>0.931</td>
<td>0.522</td>
<td>0.491</td>
<td>0.805</td>
<td>0.938</td>
<td>0.783</td>
<td>0.563</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1990</td>
<td>0.994</td>
<td>0.966</td>
<td>0.516</td>
<td>0.562</td>
<td>0.878</td>
<td>0.923</td>
<td>0.793</td>
<td>0.785</td>
<td>0.225</td>
</tr>
<tr>
<td></td>
<td>2000</td>
<td>1.016</td>
<td>0.974</td>
<td>0.590</td>
<td>0.594</td>
<td>0.930</td>
<td>0.929</td>
<td>0.813</td>
<td>0.889</td>
<td>0.310</td>
</tr>
</tbody>
</table>
It can be seen from the analysis above that the information entropy of water use can be used as a convenient measure to characterize the diversity of water use among different sectors. With the help of additional information such as the real number (other than percentages) of water use, characterization tends to be more useful in terms of comparing water use structures across regions under the same framework. Such a need is, in fact, highlighted in this study. In the Shiyang river basin, the information entropy of water use has remained at a low level but has kept growing gradually in recent years; in the Guanzhong region, values have been higher. Results indicate that the distribution of agricultural water, industrial water and municipal water are more imbalanced in the Shiyang river basin than in the Guanzhong region, due to the higher proportion of agricultural water use in the Shiyang river basin.

3.2. Analysis of the influencing factors

3.2.1. The factors influencing regional water use. The influencing factors were ranked using the MI and the determination coefficient, and are listed in Figure 3. For the MI ranking, for the Shiyang river basin the top four variables were found to be agricultural output, GDP, the effective irrigation area and industrial output, whilst for the Guanzhong region they were population, urbanization, cultivated land and grain output. When ranked by the determination coefficient (R²), the top four variables affecting the information entropy of water use, were (for the Shiyang river basin) GDP, industrial output, agricultural output and urbanization, and (for the Guanzhong region) cultivated land, industrial output, population and urbanization.

Table 3 shows that agricultural output had the highest MI value in the Shiyang river basin, and population and urbanization had higher MI values in the Guanzhong region. GDP and industrial output had higher R² in the Shiyang river basin, whilst cultivated land had the highest R² in the Guanzhong region.

Interestingly, the mutual information method tends to recognize (in terms of ranking) that agricultural water use related factors cause more change to the entropy information of water use in the Shiyang river basin, whereas factors related to industrial water use are thought to have brought more changes to the Guanzhong region. The correlation coefficient method gives different rankings of these two groups of factors for the two regions.

3.2.2. Factors influencing the different water use in the two regions. According to the results in section 3.1, there were greater differences in information entropy of water use between the Shiyang river basin and the Guanzhong region. The influencing factors were analyzed using the same method as for the regional water use (see section 3.2.1). The relationship between the differences in information entropy of water use and the differences in potential influencing factors was quantified by using the correlation coefficient and mutual information methods.

As shown in Figure 3, grain output, temperature, reference evapotranspiration (ET₀) and agricultural output were the top four variables affecting the difference in the information entropy of water use between the two regions. Both the correlation coefficient and the MI method selected five variables, i.e., grain output, agricultural output, precipitation, temperature and ET₀. According to the MI method, precipitation and ET₀ had highest values amongst all variables. In the correlation coefficient method, grain output had the highest R² value and was also the only variable that had significant correlation with the difference in the information entropy of water use. Moreover, there were significant correlations between grain output and agricultural output (R = 0.867), precipitation (R = 0.737), and ET₀ (R = 0.913).
It can be seen that the fraction of agricultural output as part of GDP was always higher in the Shiyang basin than in the Guanzhong region, reflecting the fact that socio-economic development has been more heavily dependent on agriculture. Comparing the relevant variables between the two regions in 2005, we see that there were great differences in cultivated land per person (Shiyang river basin: 2.63 ha per capita; Guanzhong region: 0.68 ha per capita), agricultural water use per person (Shiyang river basin: 1082 m³ per capita; Guanzhong region: 130 m³ per capita) and total water use per person (Shiyang river basin: 1241 m³ per capita; Guanzhong region: 211 m³ per capita), indicating that there were greater agricultural areas and agricultural water use in the Shiyang river basin. Considering the meteorological factors (lower precipitation and higher evaporation), the irrigation per area is higher in Shiyang river basin than that in the Guanzhong region. Thus the difference of water use between the two regions

Fig. 3. Determination coefficient ($R^2$) and mutual information (MI) between potential influencing factors and the information entropy of water use, showing regional differences between the Guanzhong region and the Shiyang river basin in water use sectors. (The dashed line denotes the 0.05 significant level in $R^2$).
can be attributed to both agricultural variables (grain output and agricultural output) and meteorological variables (temperature, precipitation and ET0).

3.3. Simulation of the information entropy of regional water use

3.3.1. Model performance. Table 2 shows that the model built using the MI method has the lowest root mean square error (RMSE) for both regions, which may be due to the fact that the MI method could combine the linear and nonlinear relationship between the influencing variables and the regional water use proportion. The model built using the LOO method of cross validation (CV) has a lower RMSE for the testing samples than that using the three-fold CV method. Generally, more samples give more explanations about the relationship changes between the selected variables and the information entropy of water use. Table 2 also shows that the RMSE of the three models for the Guanzhong region varied slightly more than it did for the Shiyang river basin. The main purpose of using various cross validation methods is to ensure the robustness of the models that have been built.

Table 2. Average root mean square error (RMSE) of multiple-linear regression models from different input methods for information entropy of water use in the Guanzhong region and the Shiyang river basin. The R_model takes four variables with highest correlation coefficients (R), the MI_model chooses four variables with highest mutation information values, and the Full_model utilizes all 11 variables as inputs. Two kinds of cross validation (CV) are presented for model comparison: three-fold (CV) (K = 3) and leave-one-out (LOO) CV.

<table>
<thead>
<tr>
<th>Region</th>
<th>Model</th>
<th>RMSE 3-fold CV</th>
<th>LOO CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shiyang river basin</td>
<td>R_model</td>
<td>0.131</td>
<td>0.048</td>
</tr>
<tr>
<td></td>
<td>MI_model</td>
<td>0.113</td>
<td>0.032</td>
</tr>
<tr>
<td></td>
<td>Full_model</td>
<td>0.271</td>
<td>0.128</td>
</tr>
<tr>
<td>Guanzhong region</td>
<td>R_model</td>
<td>0.114</td>
<td>0.02</td>
</tr>
<tr>
<td></td>
<td>MI_model</td>
<td>0.052</td>
<td>0.019</td>
</tr>
<tr>
<td></td>
<td>Full_model</td>
<td>0.073</td>
<td>0.02</td>
</tr>
</tbody>
</table>

Table 3. MI_model for the information entropy of regional water use in the Guanzhong region and the Shiyang river basin. The MI_model chooses four variables with highest mutation information values: R for the performance of simulation; F-test for the overall significance; t-test for the individual significance. $F_{0.05} (4, 7) = 4.12$, $t_{0.05} (12) = 1.7823$, $R_{0.05} = 0.5760$ and $R_{0.01} = 0.7079$.

<table>
<thead>
<tr>
<th>Region</th>
<th>R or F</th>
<th>Parameter</th>
<th>Value</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shiyang river basin</td>
<td>$R = 0.926$, $F = 10.51$</td>
<td>Constant ($\beta_{0,\text{MI}}$)</td>
<td>$-0.251$</td>
<td>$-1.252$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Agricultural output ($\beta_{1,\text{MI}}$)</td>
<td>$-0.013$</td>
<td>$-2.392$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gross domestic product ($\beta_{2,\text{MI}}$)</td>
<td>0.002</td>
<td>2.555</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Irrigation area ($\beta_{3,\text{MI}}$)</td>
<td>0.002</td>
<td>2.672</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Industrial output ($\beta_{4,\text{MI}}$)</td>
<td>0.001</td>
<td>1.125</td>
</tr>
<tr>
<td>Guanzhong region</td>
<td>$R = 0.937$, $F = 12.56$</td>
<td>Constant ($\beta_{0,\text{MI}}$)</td>
<td>4.243</td>
<td>2.276</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Population ($\beta_{1,\text{MI}}$)</td>
<td>0.002</td>
<td>1.963</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Urban population ($\beta_{2,\text{MI}}$)</td>
<td>$-21.191$</td>
<td>$-3.479$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cultivated land ($\beta_{3,\text{MI}}$)</td>
<td>$-0.002$</td>
<td>$-4.294$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Industrial output ($\beta_{4,\text{MI}}$)</td>
<td>0.000</td>
<td>2.467</td>
</tr>
</tbody>
</table>

can be attributed to both agricultural variables (grain output and agricultural output) and meteorological variables (temperature, precipitation and ET0).

3.3. Simulation of the information entropy of regional water use

3.3.1. Model performance. Table 2 shows that the model built using the MI method has the lowest root mean square error (RMSE) for both regions, which may be due to the fact that the MI method could combine the linear and nonlinear relationship between the influencing variables and the regional water use proportion. The model built using the LOO method of cross validation (CV) has a lower RMSE for the testing samples than that using the three-fold CV method. Generally, more samples give more explanations about the relationship changes between the selected variables and the information entropy of water use. Table 2 also shows that the RMSE of the three models for the Guanzhong region varied slightly more than it did for the Shiyang river basin. The main purpose of using various cross validation methods is to ensure the robustness of the models that have been built.
from the very limited number of samples. For the Guanzhong region, the models behave less sensitively regarding the use of different cross validation methods than do those for the Shiyang river basin. The reason for this could be that the relationship of the former region is robust enough so that it works for the entire samples while several regimes may exist for the Shiyang river basin, in which case a proper cross validation method like LOO CV can address the problem of inappropriate sampling.

3.3.2. Model simulation. According to the RMSE values obtained in the cross validation method, the MI_model was the best model in simulating the information entropy of water use. Results of the two MI_models built on the entire dataset (1994–2005) are shown in Table 2. Both models selected four input variables as influencing factors whose overall significance was validated using the F-test. The individual significance of each influencing factor was also tested using the t-test. Table 3 shows that the models of the information entropy of regional water use in both regions had overall significance; the t-test gives their individual significance. The coefficients of all variables in the Guanzhong region and the coefficients of agricultural output, gross domestic product and irrigation area in the Shiyang river basin had individual significance.

It can be seen that the information entropy of water use in the Shiyang river basin is negatively correlated with agricultural output but positively correlated with effective irrigation area, industrial output and GDP. In fact, when the irrigation area decreased sharply due to governmental control, and the industrial output and GDP increased greatly from 2005 to 2008 (see Figure 2), there was an accompanying increase in information entropy. For the Guanzhong region, the information entropy was positively related to population and industrial output and negatively related to urbanization and cultivated land, as seen by the coefficients of those influencing variables listed in Table 3.

4. Conclusions

Due to the higher proportion of agriculture water in the Shiyang river basin, the information entropy of water use was lower than in the Guanzhong region. The related variables to agricultural development (grain output, agricultural output) and meteorological factors (temperature, precipitation and reference evapotranspiration) were the main variables causing the difference of regional water use in both regions. The mutual information (MI) and correlation coefficients were used to help build an efficient multiple-linear regression (MLR) model analyzing the complex relationship between regional water use and influencing factors. It has been shown that the MLR model with influencing factors chosen using an MI score is the best model in terms of simulating changes in regional water use. As indicated by the root mean square error (RMSE) from cross validation, the relationship between influencing factors and regional water use tended to be more stable in Guanzhong than in the Shiyang river basin. The information entropy of water use represents the change of water use proportion and can be used for regional comparison; the MI score is shown to be very helpful in building MLRs for modeling water use with very limited data available.

Acknowledgements

We are grateful for the grant support from the China Scholarship Council of China, UNESCO-IHE of the Netherlands, and the research grants from the Ministry of Science and Technology of China (2011AA100502) and from the Ministry of Water Resources of China (201201003).
References


Received 18 April 2011; accepted in revised form 25 January 2012. Available online 4 June 2012.