

Forecasting monthly energy production of small hydropower plants in ungauged basins using grey model and improved seasonal index

Chun-Tian Cheng, Shu-Min Miao, Bin Luo and Yong-Jun Sun

ABSTRACT

A first-order one-variable grey model (GM(1,1)) is combined with improved seasonal index (ISI) to forecast monthly energy production for small hydropower plants (SHPs) in an ungauged basin, in which the ISI is used to weaken the seasonality of input data for the GM(1,1) model. The ISI is calculated by a hybrid model combining K-means clustering technique and ratio-to-moving-average method, which can adapt to different inflow scenarios. Based on the similar hydrological and meteorological conditions of large hydropower plants (LHPs) and SHPs in the same basin, a reference LHP is identified and its local inflow data, instead of the limited available data of SHPs, is used to calculate the ISI. Case study results for the Yangbi and Yingjiang counties in Yunnan Province, China are evaluated against observed data. Compared with the original GM(1,1) model, the GM(1,1) model combined with traditional seasonal index (TSI-GM(1,1)), and the linear regression model, the proposed ISI-GM(1,1) model gives the best performance, suggesting that it is a feasible way to forecast monthly energy production for SHPs in data-sparse areas.

Key words | grey model, improved seasonal index, monthly energy production forecasting, small hydropower plant, ungauged basins

Chun-Tian Cheng (corresponding author)

Shu-Min Miao

Bin Luo

Yong-Jun Sun

Institute of Hydropower System &

Hydroinformatics,

Dalian University of Technology,

Dalian 116024,

China

E-mail: ctcheng@dlut.edu.cn

Chun-Tian Cheng

Key Laboratory of Ocean Energy Utilization and

Energy Conservation of Ministry of Education,

Dalian University of Technology,

Dalian 116024,

China

Bin Luo

Tsinghua Sichuan Energy Internet Research

Institute,

Chengdu 610200,

China

Yong-Jun Sun

Kunming Power Exchange Center,

Kunming 650200,

China

INTRODUCTION

Along with rapid social-economic development, growing environmental degradation and climate change, renewable energy becomes increasingly important in the energy supply portfolio (Panwar *et al.* 2011; REN21 2014). As a favorable energy source of the Clean Development Mechanism and a typical renewable energy, small hydropower contributes to low-carbon and sustainable development, and has attracted renewed interest worldwide (Turkey: Bakis & Demirbas 2004; India: Dudhani *et al.* 2006; Purohit 2008; Africa: Taele *et al.* 2012; Kaunda 2013; China: Hong *et al.* 2013; Cheng *et al.* 2015; Brazil: Ferreira *et al.* 2016). China has the world's biggest installed capacity for small hydropower with more than 75 million kilowatt by the end of 2015, and has become a crucial component of the national electricity supply (Kong *et al.* 2015). With more small hydropower

plants (SHPs) being integrated into power systems, the difficulty in coordinated dispatching between SHPs and other power sources has significantly increased, causing more serious transmission congestion and spilled water problems. The inherent variability of SHP generation also poses a direct threat to the security and reliability of power systems. Thus, an accurate forecasting for long-term SHP energy production is essential for power systems' operation and dispatching. Nevertheless, this is not an easy task, as most SHPs are located in remote areas with few hydrological and meteorological stations and lack of professional supervision for a long time, their historical information is absent.

Regarding the published forecasting models relating to hydropower matters, most have been for large hydropower plants (LHPs) and mainly focused on forecasting of river

flows, such as stream flow (Kim & Seo 2015; Li *et al.* 2015c; Taormina & Chau 2015), reservoir inflow (Valipour *et al.* 2013; Bai *et al.* 2016), or rainfall and runoff (Chau & Wu 2010; Wang *et al.* 2015). Only a few studies have been carried out on forecasting of small hydropower production, and have predominantly focused on the short-term horizon (Esteperez & Nagasaka 2006; Monteiro *et al.* 2013, 2014; Li *et al.* 2015a, 2015b). Basically, these published forecasting models can be divided into four categories: physical models (Golmar *et al.* 2017), statistical models (Taormina & Chau 2015), artificial intelligence models (Li *et al.* 2015b), and hybrid models (Bai *et al.* 2016). The input data and model parameter sensitivity analysis and uncertainty estimation methods have also been widely studied (Srivastava *et al.* 2014; Tong *et al.* 2016; Tongal & Booi 2017). However, most of these existing models typically require large numbers of historical observations or complicated input variables, such as reservoir inflow, atmospheric temperature, precipitation, among others, which are what SHP in ungauged basins lack.

The Grey System Theory (GST) (Deng 1982) provides an alternative solution, with the main focus on modeling with small data sets and imperfect information. As the main forecasting model in GST, the grey model (GM(1,1)) model has been applied in various forecasting problems (Yao *et al.* 2003; Alvisi *et al.* 2013; Yin 2013; Xie *et al.* 2015), but its application in small hydropower problems is rare and still to be investigated. The difficulties or disadvantages of the original GM(1,1) model in forecasting SHP energy production include the following:

1. Due to the lack of a regulation reservoir, the SHP power generation obviously fluctuates with the variation of seasons, which leads to poor results for the GM(1,1) model with original energy production data (Deng 1989).
2. Seasonal index is a feasible way to eliminate the seasonal variation of energy production (Taylor 2010); however with a small data set of SHP, an effective seasonal index is difficult to construct.
3. The number of SHPs is numerous, and the forecasting workload would certainly be heavy when predicting for each of them.

To overcome these problems and achieve a successful implementation for monthly energy production forecasting for SHP with few available data, the GM(1,1) model

combined with an improved seasonal index (ISI-GM(1,1)) is proposed in this paper. The main contributions of this work can be summarized as follows:

1. The SHPs located in the same region are treated as a group, so as to weaken the stochastic fluctuations of individual ones and also reduce the prediction workload. This is reasonable since they share similar hydrological and meteorological conditions, and generally exert a group influence on power system operation. In what follows, the SHPs mean those located in the same region.
2. An ISI is proposed. Compared with the traditional constant seasonal index, the ISI is more suitable for different inflow scenarios (i.e., wet, normal, and dry), so as to better weaken the seasonality of energy production data sequence. A hybrid model combining K-means clustering technique and ratio-to-moving-average (RMA) method is also developed for calculating the ISI.
3. The correlation between LHP local inflow and SHPs' energy production is noted and carefully analyzed, the LHP showing significant correlation and with sufficient data series is selected as the reference LHP. The local inflow of this reference LHP, instead of the limited SHPs data, is thus applied to construct the ISI.
4. The input data for the GM(1,1) model is the processed SHPs' energy production data by using the calculated ISI, not the raw data sequence, so as to improve the forecasting accuracy.
5. The forecasting performance of the proposed model has been evaluated by applying it for forecasting SHPs' monthly energy production of Yangbi and Yingjiang counties in Yunnan province, China. Further comparisons between the proposed model and other models, including the original GM(1,1) model, the GM(1,1) model combined with traditional seasonal index (TSI-GM(1,1)) and the linear regression model (LR) from a previous study (Li *et al.* 2015a), are also discussed. The results show that the proposed ISI-GM(1,1) model is a feasible way for monthly energy production forecasting for SHPs in ungauged basins.

The following sections contain a description of the proposed forecasting method, in which the ISI, the GM(1,1) model, and the ISI-GM(1,1) model are introduced. The listing of the performance evaluation criteria used in this paper and

a brief introduction for the study areas are also presented. Then, the simulation results made by the proposed model for the actual case study are given and also compared with other models. Finally, conclusions are drawn.

METHODOLOGY

Preparation for modeling

As mentioned above, the individual SHP energy production has strong uncertainty due to the lack of a regulation reservoir and is directly affected by the stochastic natural water inflow. Thus, it is difficult to obtain the energy production rules and develop a suitable forecasting model for individual SHPs. Furthermore, even if a forecasting model can be established for an individual SHP, the forecasting workload would be very large given the large number of SHPs. Therefore, in this paper, the SHPs located in the same region are taken as a group to build the forecasting model. In this way, the potential energy production rules can be more easily acquired, and the forecasting workload is significantly reduced. This is reasonable, because the SHPs located in the same region are generally integrated into the power grid via the same transmission line, and exert a group influence on power system operation. To eliminate the influence of different installed capacity on energy production, the monthly utilization hours are introduced, i.e.:

$$h_k = \frac{g_k}{c_k}, k = 1, 2, \dots, n \quad (1)$$

where h_k is the monthly utilization hours of SHPs at time k ; n is the number of time periods and each time period represents a month; g_k and c_k represent the monthly energy production and the installed capacity of SHPs at period k , respectively. Despite the variations of installed capacity, the monthly energy production data sequence is equivalent to the sequence of monthly utilization hours.

The ISI

The stationarity of the input data plays a key role in the forecasting accuracy of the GM (Deng 1989). Hence, the seasonal

index is introduced to weaken the seasonality for the SHPs' monthly utilization hours sequence. The seasonal index contains 12 separate values which correspondingly represent generation variations from January to December, and is generally calculated by the RMA method (Tseng *et al.* 2001). However, when it comes to monthly energy production forecasting for SHPs, some challenges have appeared, including: (1) the data length of SHPs in ungauged basin is too short to construct a reliable seasonal index; (2) different monthly inflow scenarios (i.e., wet, normal, and dry) have a significant effect on the fluctuations of SHPs' energy production, which are not considered in the traditional constant seasonal index; and (3) the trend information of past periods of the forecasting period is neglected, which is unreasonable, because the energy production is somehow contiguous to the past periods. To solve the above problems, an ISI is proposed and a hybrid model combining K-means clustering technique and RMA method is also developed to calculate it.

Selection of reference LHP

It is hard to construct a reliable ISI using limited historical data of SHPs. Generally, the SHP is built on a small river without a regulation reservoir, and its energy production is mainly determined by the natural water inflow. As there are similar hydrological and meteorological conditions of SHPs and LHPs in the same basin, the LHP local inflow (contribution from the sub-basin in the reservoir and all its immediate upper reservoirs), at a certain degree, reflects the natural water inflow of SHPs. Thus, a reference LHP is identified and its sufficient local inflow data are used to calculate the ISI. The procedures for selecting the reference LHP are summarized below:

1. Selection of candidate LHPs: Due to the lack of accurate locations of SHPs, the reference LHP cannot be directly identified by observation. Thus, for better results, all LHPs in the same region are considered as candidates. It should be noted that, in this paper, the LHP includes LHPs as well as the downstream hydropower plants. That is to say, the downstream hydropower plant with small installed capacity is not included in the studied SHPs, because these plants, located on main rivers with LHP, are also supervised well. This paper mainly focuses

on the SHPs with limited available data, whose energy production cannot be directly calculated.

- Correlation analysis and significance test: The correlation coefficient between the LHP local inflow and the utilization hours of SHPs is calculated and evaluated by significance test. As is well known, the correlation coefficient is a function that is commonly used to indicate the degree of correlations between two sets of observed data (Zhu & Yuan 2015), and can be calculated as follows:

$$R = \frac{\sum_{k=1}^n (h_k - \bar{h})(q_k - \bar{q})}{\sqrt{\sum_{k=1}^n (h_k - \bar{h})^2 \sum_{k=1}^n (q_k - \bar{q})^2}} \quad (2)$$

where \bar{h} is the mean value for SHPs monthly utilization hours from n periods; q_k is the local inflow of reference LHP at period k ; \bar{q} is the mean local inflow value of reference LHP from n periods; R is the correlation coefficient. Based on the results, the LHPs which show positive and significant correlations with SHPs are selected. The significance test has been described in detail in Li et al. (2015a) and is not repeated here.

- Identification of reference LHP: The LHP not only shows significant correlation, but also has sufficient data series, and is finally selected as the reference LHP as adequate data are necessary to construct a reliable seasonal index.

Calculation of ISI

In the hybrid model, the K-means clustering technique is introduced to divide reference LHP local inflow data into subsets which can be analyzed separately. K-means finds the homogeneous groups for original data points by minimizing the sum of squared error between each data point and the closest centroid (Tan et al. 2014). The calculating process of ISI is shown in Figure 1 and detailed below:

- Step 1:** Define the monthly reservoir local inflow of reference LHP as $\{q_{i,t}, i = 1, 2, \dots, m; t = 1, 2, \dots, 12\}$, here, i is the year index; m is the number of years; t is the month index, where $t = 1$ represents January, $t = 2$ represents February, ..., and $t = 12$ represents December.
- Step 2:** The $\{q_{i,t}, i = 1, 2, \dots, m; t = 1, 2, \dots, 12\}$ is clustered separately for each month (i.e., January, February, ...,

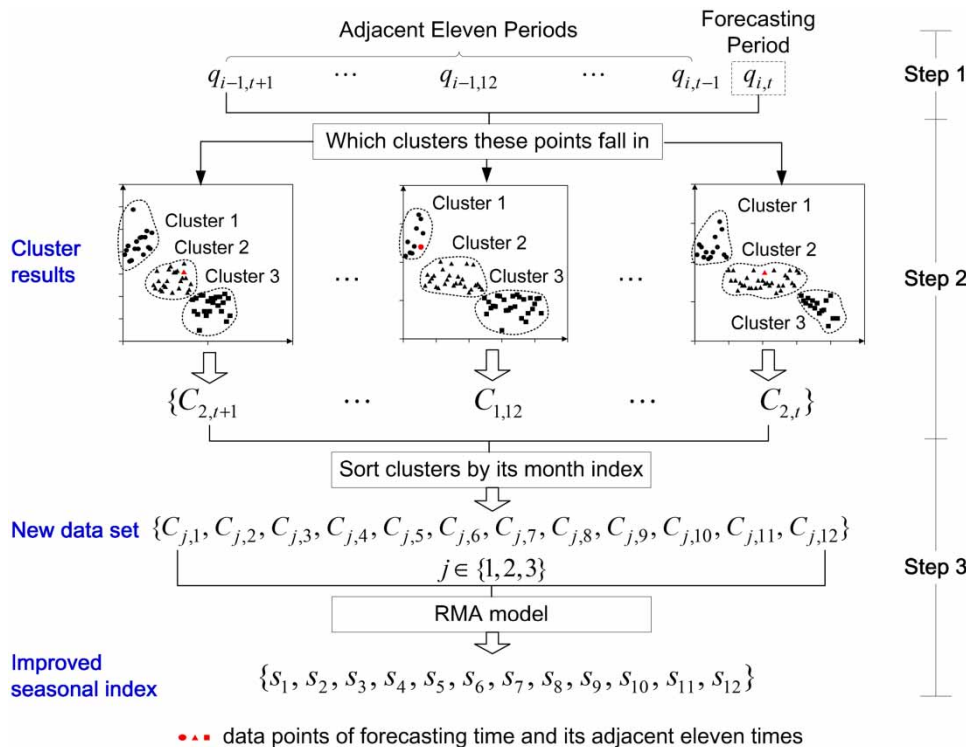


Figure 1 | The calculating process of ISI.

December) by K-means. The data of each month is divided into three subsets which may be associated with dry-, normal- and wet-inflow. These subsets are expressed as $\{C_{j,t}, j = 1, 2, 3\}$, where $C_{j,t}$ is the subset of month t ; $j = 1$ represents dry scenario, $j = 2$ represents normal scenario and $j = 3$ represents wet scenario. For example, $C_{3,1}$ is composed of all January observations which fall in the subset of wet scenario. The implementation steps of K-means are described in detail in Tan et al. (2014) and are not repeated here.

Step 3: Suppose that the forecasting period is month t of year i , the subsets which the forecasting period and its adjacent eleven periods fall in are obtained by Step (2). These subsets are ordered by their month index, and form a new data set. The ISI of each month is calculated by the RMA model from this new data set. The calculation procedures of the RMA model are reported in detail in Tseng et al. (2001).

When forecasting for another time period, the ISI should be recalculated. It is important to note that the cluster number used in the proposed model is three, which is determined by considering the traditional classification of inflow scenarios, i.e., dry, normal, and wet. The cluster number can be set to other values and the forecasting performances with different cluster numbers are compared and analyzed in the section 'Discussion on cluster number'.

Combining ISI with the first-order one-variable GM

The first-order one-variable grey model (GM(1,1))

The GM(1,1) model is one of the most frequently used grey forecasting models, which is characterized by modeling with few available data and ease of calculation. To enhance forecasting accuracy, it develops two necessary data operations: one is the accumulated generating operation (AGO), and the other is the inverse accumulative generating operation (IAGO). The AGO aims to weaken the fluctuations of original data series, and the IAGO is used to recover the AGO generated data to the original data sequence. The steps implemented in the GM(1,1) model are listed below:

Step 1: Define the original data for GM(1,1) model as $\{x^{(0)}(k), k = 1, 2, \dots, n\}$, and the time period to be forecast

is $n + 1$. In this paper, $x^{(0)}(k) = h_k/s_k$, where k is the period number, $k = 1, 2, \dots, n$; h_k and s_k are the monthly utilization hours and the ISI of period k , respectively.

Step 2: A new sequence $x^{(1)}(k)$, ($k = 1, 2, \dots, n$) generated by AGO from $x^{(0)}(k)$ is derived as Equation (3):

$$x^{(1)}(k) = \sum_{i=1}^k x^{(0)}(i), (k = 1, 2, \dots, n) \quad (3)$$

Then, the first order differential equation can be formed by using $x^{(1)}$, expressed as Equation (4):

$$\frac{dx^{(1)}}{dt} + ax^{(1)} = u \quad (4)$$

where k is the time period number, a and u are the optimization parameters.

Step 3: By using the least squares method, a and u can be determined:

$$\hat{a} = [a, u]^T = (B^T B)^{-1} B^T y_N \quad (5)$$

where

$$B = \begin{pmatrix} -\frac{1}{2}[x^{(1)}(1) + x^{(1)}(2)] & 1 \\ -\frac{1}{2}[x^{(1)}(2) + x^{(1)}(3)] & 1 \\ \dots & \dots \\ -\frac{1}{2}[x^{(1)}(n-1) + x^{(1)}(n)] & 1 \end{pmatrix},$$

$y_N = [x^{(0)}(2), x^{(0)}(3), \dots, x^{(0)}(n)]^T$ and B^T is the transpose matrix of B .

Step 4: Based on the above steps, the simulating value of $x^{(1)}(k + 1)$ is calculated by Equation (6):

$$\hat{x}^{(1)}(k + 1) = \left[x^{(0)}(1) - \frac{u}{a} \right] e^{-ak} + \frac{u}{a} \quad (6)$$

where $\hat{x}^{(1)}(k + 1)$ is the simulating value of $x^{(1)}(k + 1)$.

Step 5: Then, the forecasting value of time $n + 1$ is obtained by IAGO from simulating values $\{\hat{x}^{(1)}(k + 1), k = 1, 2, \dots, n\}$, expressed as Equation (7):

$$\hat{x}^{(0)}(n + 1) = \hat{x}^{(1)}(n + 1) - \hat{x}^{(1)}(n) \quad (7)$$

where $\hat{x}^{(0)}(n + 1)$ is the forecasting value of period $n + 1$.

The ISI-GM(1,1) model

Due to the natural seasonal fluctuations of SHPs’ energy production, the ISI is combined with the GM(1,1) model. The forecasting process of the ISI-GM(1,1) model for SHPs’ monthly utilization hours are described as follows:

Step 1: Suppose that the forecasting time period is $n + 1$. Identify the adjacent 11 periods of the forecasting period, and construct the ISI from the reference LHP local inflow data of these 12 periods.

Step 2: The data of SHPs’ monthly utilization hours are divided by the calculated ISI, then a new data series is obtained, expressed as Equation (8):

$$h'_k = \frac{h_k}{s_t} \tag{8}$$

where h'_k is the new data series; k is the period number, $k = 1, 2, \dots, n$; h_k and s_k , respectively, represent the monthly utilization hours and the ISI of period k .

Step 3: Take the new data series h'_k as input data for the GM (1,1) model, that is, $x^{(0)}(k) = h'_k, k = 1, 2, \dots, n$. Then, the simulated value $\hat{x}^{(0)}(n + 1)$ is obtained by the steps outlined in the section ‘The first-order one-variable GM’.

Step 4: The forecasting value of SHPs’ monthly utilization hours at period $n + 1$ is obtained through Equation (9):

$$\hat{h}_{n+1} = \hat{x}^{(0)}(n + 1) \times s \tag{9}$$

where \hat{h}_{n+1} and s are the forecasting value and the corresponding ISI of period $n + 1$, respectively.

A flow chart of the entire process of the proposed ISI-GM(1,1) model is shown in Figure 2.

FORECASTING PERFORMANCE CRITERIA

Checking method for GMS

For evaluating the fitting precision of GMS, the most widely used method is the after-test residue checking method (Cheng et al. 2016). Two main parameters, posterior-error (C) and micro-error-probability (P), are adopted, respectively defined as:

$$C = \frac{S_2}{S_1} \tag{10}$$

$$P = \text{probability}\{|\varepsilon_k - \bar{\varepsilon}| < 0.6745S_1\} \tag{11}$$

where

$$\begin{cases} S_1 = \sqrt{\frac{1}{n-1} \sum_{k=1}^n [h_k - \bar{h}]^2} \\ S_2 = \sqrt{\frac{1}{n-1} \sum_{k=1}^n [\varepsilon_k - \bar{\varepsilon}]^2} \\ \bar{h} = \frac{1}{n} \sum_{k=1}^n h_k \\ \varepsilon_k = h_k - \hat{h}_k \\ \bar{\varepsilon} = \frac{1}{n} \sum_{k=1}^n \varepsilon_k \end{cases} \tag{12}$$

The fitting precision grade is shown in Table 1.

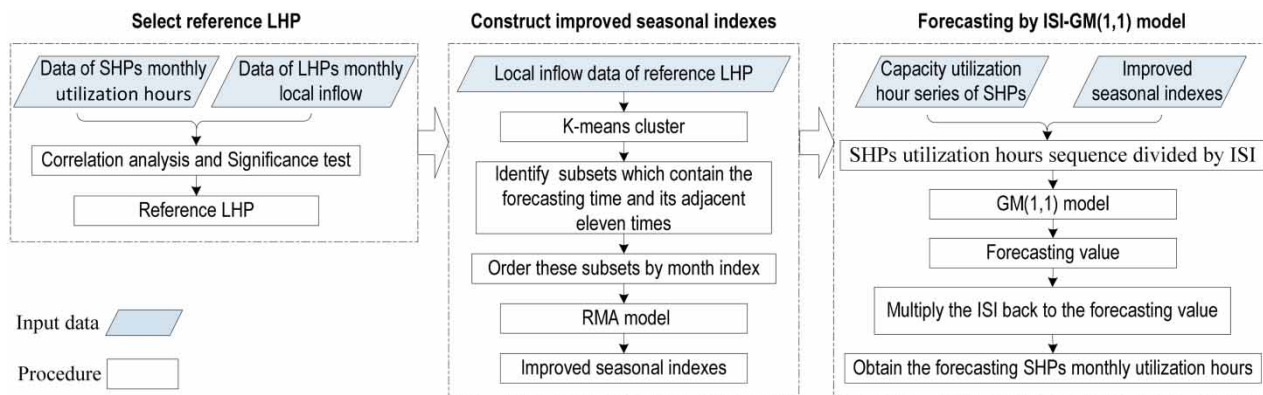


Figure 2 | Flow chart of the ISI-GM(1,1) model for monthly utilization hours forecasting.

Table 1 | Reference for fitting precision grade

Parameters	Fitting precision grade			
	Good	Qualified	Just	Unqualified
<i>C</i>	<0.35	0.35–0.50	0.50–0.65	≥0.65
<i>P</i>	>0.95	0.80–0.95	0.70–0.80	≤0.70

Performance evaluation for forecasting models

Some criteria are recommended for evaluating forecasting models according to the published literature. In this paper, four criteria are used, and computed as follows.

Root-mean-square error

The root-mean-square error (RMSE) is used to measure the difference between the values forecasted by a model and the observed values from practice, and is one of the frequently used criteria. It is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{k=1}^n (h_k - \hat{h}_k)^2} \quad (13)$$

Mean absolute percentage error

The mean absolute percentage error (MAPE) aims to evaluate the forecasting accuracy from a term-by-term comparison of the relative error in the forecasting with respect to the observed value. It usually expresses the accuracy as a percentage, and is defined as:

$$\text{MAPE} = \frac{1}{n} \sum_{k=1}^n \left| \frac{h_k - \hat{h}_k}{h_k} \right| \times 100\% \quad (14)$$

Mean absolute error

As the name suggests, the mean absolute error (MAE) is an average of the absolute errors. It is a quantity usually used to describe how close the forecasting values are to the observed values, which is a common measure of forecasting

error. It is defined as:

$$\text{MAE} = \frac{1}{n} \sum_{k=1}^n |h_k - \hat{h}_k| \quad (15)$$

Coefficient of determination

The coefficient of determination (R^2) is used to describe the degree of collinearity between the forecasting and observed values, which ranges from 0 to 1 and is defined as:

$$R^2 = \frac{\left[\sum_{k=1}^n (h_k - \bar{h})(\hat{h}_k - \bar{\hat{h}}) \right]^2}{\sum_{k=1}^n (h_k - \bar{h})^2 * \sum_{k=1}^n (\hat{h}_k - \bar{\hat{h}})^2}, \quad \bar{h} = \frac{1}{n} \sum_{k=1}^n h_k, \quad \bar{\hat{h}} = \frac{1}{n} \sum_{k=1}^n \hat{h}_k \quad (16)$$

In the above equations, n is the number of forecasting time periods; h_k and \hat{h}_k are, respectively, the observed and the forecasting value of period k . In evaluating forecasting performance, the smaller *RMSE*, *MAPE*, *MAE* and the larger R^2 indicate the better forecasting performance.

STUDY AREA AND DATA

Study areas

To ensure the similarity and transmission integrity of SHPs, a county is treated as a study unit. Two counties, Yangbi County in Dali City and Yingjiang County in Dehong City, in Yunnan Province (China) were selected as illustrating examples to demonstrate the effectiveness of the proposed ISI-GM(1,1) model. The locations of Yunnan Province and the two counties are shown in Figure 3. Yunnan Province is located in southwestern China and is extremely rich in hydropower resources, with three of China's thirteen hydro-power bases built here. By the end of 2015, the SHP installed capacity of Yunnan Province reached 10,740.5 MW, becoming the third largest provincial power resource. According to the statistics, about 3.7% and 13%

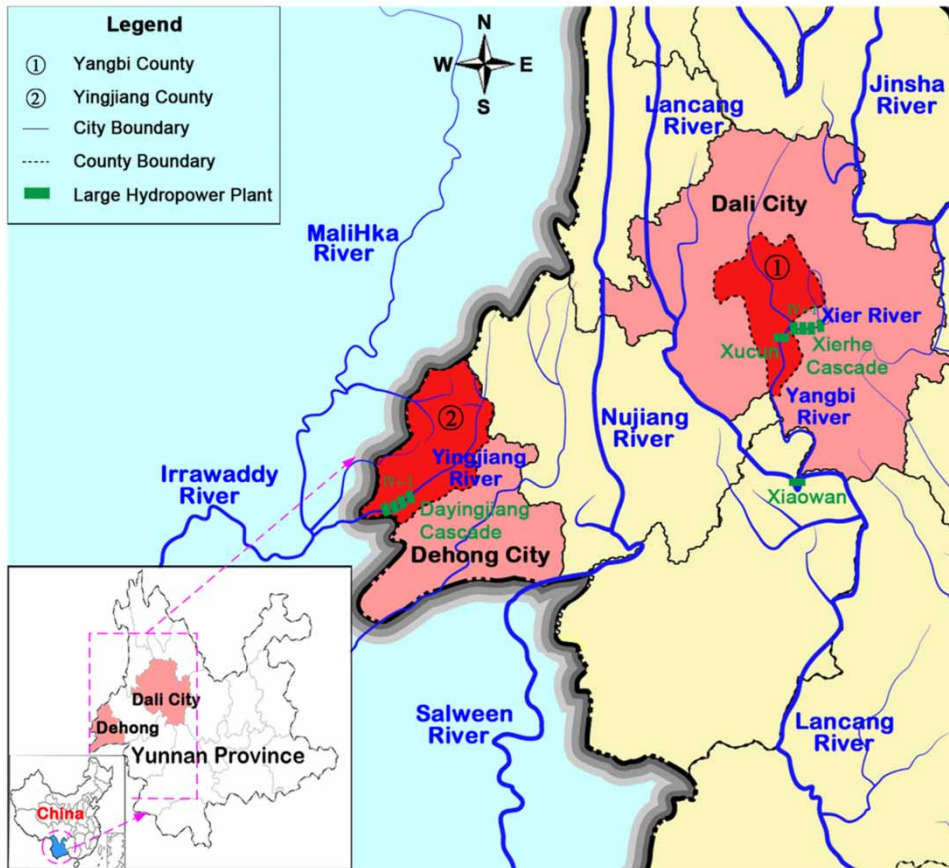


Figure 3 | Locations of Yangbi County and Yingjiang County.

of the total SHP installed capacity come from Dali City and Dehong City, respectively. The two counties studied in this article both have the richest small hydropower resources in their own city. The available information of SHPs only includes the dispatching department (i.e., county dispatching bureau), installed capacity, and the energy production data of four years (from 2012 to 2015). Thus, the accurate location for each SHP and the small tributaries they are situated on are not given in Figure 3, since this information is absent. In addition, what this paper is mostly concerned

with is the SHPs' overall impact on the power grid operation, thus the accurate location of each SHP has no or little effect on the results. Table 2 gives detailed information of these two counties.

Data collection

During the forecasting process, the observed data, including monthly energy production and installed capacity of SHPs in Yangbi and Yingjiang counties, in four years from 2012 to

Table 2 | The detailed information of the two counties

Study region	Location information		Rivers in county		SHPs in county	
	City	River system	Numbers	Largest river	Numbers	Installed capacity (MW)
Yangbi	South of Dali City	Lancang River	117	Yangbi River	28	85.52
Yingjiang	Northwest of Dehong City	Irrawaddy Basin	43	Yingjiang River	75	1,176.72

2015 are used. These observed data were collected by SHPs' operators and already validated by the Yunnan Power Grid. The local inflow data of neighboring LHPs, which are used to construct ISI, are also used. The local inflow is determined by the natural inflow from the sub-basin in the LHP reservoir and all its immediate upper reservoirs, and is not influenced by the stored/released water from upstream reservoir.

APPLICATIONS

Reference LHP

To construct an effective ISI for SHPs, a reference LHP is carefully selected, of which the local inflow is used as the input data. In Yangbi County and Yingjiang County, there are, respectively, six and four LHPs that can be considered as candidates

for reference. The detailed information of these LHPs are listed in Table 3, and their locations are given in Figure 3. According to Equation (2), the correlation coefficients between LHP local inflow and SHPs' utilization hours are calculated by using the data from January 2012 to December 2015.

In Figure 3 it can be seen that the Xier River feeds into the Yangbi River, and then to the Lancang River in the upstream of Xiaowan plant. Although the accurate locations of SHPs are not known, they must not be on the main rivers (i.e., Yangbi River, Xier River, and Yingjiang River) because the plant information for these rivers is apparent. Thus, based on the relative positions of LHP and SHP, as shown in Figure 4, there are usually two typical cases: (1) a small river with an SHP feeds into its higher-order river at the downstream of a LHP and there is no direct streamflow connection between the SHP and the LHP; and (2) a small river with an SHP feeds into its higher-order river at the upstream

Table 3 | The detailed information of candidate LHPs

Study region	LHP	Location information	Installed capacity (MW)	Correlation coefficient	Data length (year)
Yangbi	Xucun	In Yangbi County and on the mainstream of Yangbi River	84	0.95	10
	Xierhe-I	Near Yangbi County and on a tributary of Yangbi River	105	0.56	68
	Xierhe-II		50	0.64	63
	Xierhe-III		50	0.87	63
	Xierhe-IV	In Yangbi County and on a tributary of Yangbi River	50	0.88	63
	Xiaowan ^a	In the confluence of Yangbi River and Lancang River	4,200	0.94	63
Yingjiang	Dayingjiang-I ^a	In Yingjiang County and on the mainstream of Yingjiang River	108	0.91	61
	Dayingjiang-II		70	0.85	63
	Dayingjiang-III		196	0.87	63
	Dayingjiang-IV		875	0.88	10

^aThe reference LHP.

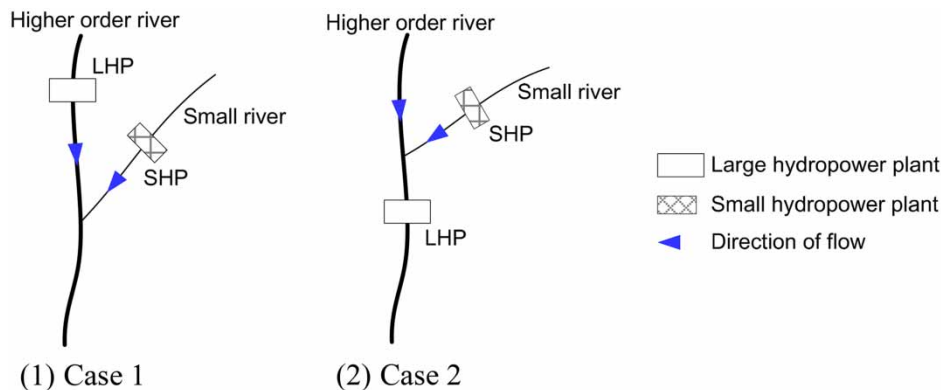


Figure 4 | The streamflow connections between LHP and SHP.

of a LHP, and the SHP inflow contributes to part of the LHP local inflow. The modified outflow of LHP also has no influence on the SHP energy production.

For Yangbi County, as shown in Figure 3, probably both the two cases exist. The SHPs located on those small rivers that feed into the Yangbi River at the downstream of Xucun plant fit Case 1, and the others fit Case 2. From Figure 3, according to the location of Xucun plant and Xierhe cascade, most of the SHPs in this county may fit Case 2, whose inflow contributes to the local inflow of Xucun and Xiaowan plants. Thus, the local inflow of Xucun and Xiaowan plants may have better correlations with the SHPs' energy production. However, due to the lack of the accurate location of each SHP, the final reference LHP should be further verified by correlation analysis and significance test. The statistical results show that all LHPs passed the significance test (at 0.01 level), and the correlation coefficients of Xucun plant and Xiaowan plant are almost the same and higher than other LHPs. However, the length of data of Xucun plant is too short to construct a reliable seasonal index while Xiaowan plant has too long a data series. Thus, Xiaowan plant was finally selected as the reference LHP. In addition, it is worth noting that the local inflow of Xiaowan plant has accepted the regulated outflow from Xucun and Xierhe cascade.

Similar to SHPs in Yingjiang County, there are no SHPs located on the Yingjiang River. Dayingjiang-I plant may show the best correlations with the energy production of SHPs in Yingjiang County because most of the SHPs are likely to be situated on the small side tributaries flowing to the Yingjiang River at the upstream of Dayingjiang-I plant, and contribute part of Dayingjiang-I's local inflow, which fits Case 1. The results of the correlation analysis and significance test show that all four LHPs have high correlations. This is because that all plants in Dayingjiang cascade are run-of-river plants, and the high correlation coefficients of Dayingjiang-II, -III, and -IV are directly influenced by the natural inflow

of Dayingjiang-I plant. Although the length of data for Dayingjiang-I plant is shorter than Dayingjiang-II and Dayingjiang-III plants, the difference is very small. Therefore, Dayingjiang-I plant was finally selected as the reference LHP.

An example of forecasting procedure

To describe the forecasting procedures of the proposed ISIGM(1,1) model in more detail, an example for August 2014 in Yangbi County is given. The input data are the SHPs' monthly energy production series from January 2012 to July 2014. The monthly local inflow data from January 1953 to August 2014 of Xiaowan plant are used to construct the ISI, in which the local inflow in August 2014 is provided by streamflow prediction software applied in practical operation of Yunnan Power Grid and is treated as known data. The proposed model was programmed via Java programming language and the used K-means clustering technique is from the WEKA Java Package (Bouckaert *et al.* 2010). The detailed procedures are described as follows:

1. The monthly local inflow data of Xiaowan plant was clustered separately for each month.
2. Twelve subsets which contain the forecasting time period (August 2014) and its adjacent 11 time periods (from July 2014 backtracking to September 2013) were obtained from 1) and ordered by month index, i.e., from January to December, as shown in Table 4. Here, $C_{1,t}$, $C_{2,t}$ and $C_{3,t}$ represent subset 1, subset 2, and subset 3 of month t , respectively, i.e., dry inflow scenario, normal inflow scenario, and wet inflow scenario.
3. Then, the ISI of each month was calculated by the RMA model, and also normalized so as to ensure the calculated accuracy in the division program, also listed in Table 4.
4. The monthly energy production was transformed into monthly utilization hours by Equation (1).

Table 4 | The inflow statuses of Xiaowan plant from September 2013 to August 2014

Time periods	Jan. 2014	Feb. 2014	Mar. 2014	Apr. 2014	May 2014	Jun. 2014	Jul. 2014	Aug. 2014	Sep. 2013	Oct. 2013	Nov. 2013	Dec. 2013
Cluster	$C_{3,1}$	$C_{3,2}$	$C_{1,3}$	$C_{1,4}$	$C_{1,5}$	$C_{1,6}$	$C_{3,7}$	$C_{1,8}$	$C_{1,9}$	$C_{1,10}$	$C_{1,11}$	$C_{3,12}$
ISI	0.659	0.563	0.456	0.575	0.384	0.779	1.079	2.41	2.028	1.618	0.863	0.587
Normalized ISI	0.055	0.047	0.038	0.048	0.032	0.065	0.090	0.201	0.169	0.135	0.072	0.049

5. A new data series was obtained by Equation (8) from the monthly utilization hours and the normalized ISI, which was used as input data for the GM(1,1) model. The obtained forecasting value was 3,247.56.
6. Finally, the forecasting value should multiply by the ISI of August, as shown in Equation (9). Hence, the forecasting utilization hours of SHPs in Yangbi County in August 2014 was $3,247.56 \times 0.201 = 652.76$ h. The error between observed value and forecasting value was $|652.76 - 597.9| \div 597.9 \times 100\% = 9.2\%$, which is acceptable.

Forecasting results of ISI-GM(1,1) model

Based on the calculated monthly utilization hours of Yangbi and Yingjiang counties during January 2014 to December 2015 and the local inflow data of reference LHPs (i.e., Xiaowan plant and Dyingjiang-I plant), the forecasting results can be achieved by the steps outlined in the section ‘The ISI-GM(1,1) model’, and shown in Figure 5.

It can be observed that the forecasting values can follow the changes of the observed data, and the probability of periods whose relative error is smaller than 10% are, respectively, 83.3% and 100.0%, in Yangbi County and Yingjiang County. The performance evaluation criteria for Yangbi County are RMSE = 38.71, MAPE = 9.93%, MAE = 23.38, and $R^2 = 0.962$, respectively. For Yingjiang County, these criteria are RMSE = 34.65, MAPE = 4.05%, MAE = 17.20, and $R^2 = 0.973$, respectively. For fitting precision checking, the *C* and *P* in Yangbi County are 0.21 and 100.0%, respectively. In Yingjiang County, the *C* and *P* are 0.20 and 95.8%,

respectively. In both the two study regions, the proposed GM gets ‘Good’ grade. These results illustrate that the proposed ISI-GM(1,1) model performs well in forecasting monthly energy production of SHPs in data-sparse areas.

Comparisons with other models

To further illustrate the forecasting performance of the proposed ISI-GM(1,1) model, three other models, i.e., the GM(1,1) model, the TSI-GM(1,1) model, and the LR model, were established under the same data set. The GM(1,1) model is characterized by its prominent capability in modeling with a small data set and imperfect information; however, its forecasting accuracy is unsatisfactory for data sequence with obvious seasonality. In addition, the TSI-GM(1,1) model is a hybrid model combining the traditional constant seasonal index and the GM(1,1) model, in which the seasonal index is directly calculated by RMA method from the not clustered local inflow data of reference LHP. Although the TSI, to a certain degree, eliminates the seasonality of input data, the constant values cannot always perform well in different inflow scenarios. The LR model is a typical time series model, with the advantages of simple modeling and easily solvable. In our previous study, this model was used to forecast the monthly energy production of SHPs in an ungauged basin and some achievements were attained (Li *et al.* 2015a). However, with gradual extensive applications, we found that this model cannot provide satisfactory results when the inflow of the next period is away from the regression curve. The

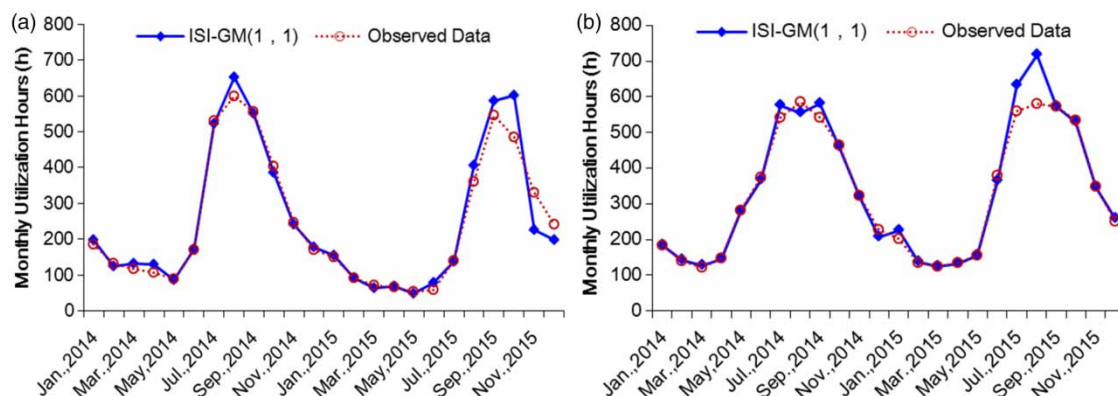


Figure 5 | Forecasting results of the proposed model for Yangbi County and Yingjiang County during January 2014 to December 2015. (a) Yangbi County. (b) Yingjiang County.

input data of the proposed ISI-GM(1,1), GM(1,1), and TSI-GM(1,1) models are, respectively, SHPs utilization hours processed by using ISI, the raw SHPs utilization hours, and SHPs utilization hours processed by using TSI. The LR model is modeled from the linear relationship between the local inflow of reference LHP and the SHPs utilization hours, for Yangbi and Yingjiang counties the fitting model are, respectively:

$$\hat{h}_{n+1} = 19.08 + 2.89 \times q_{n+1} \quad (17)$$

$$\hat{h}_{n+1} = 124.25 + 1.11 \times q_{n+1} \quad (18)$$

Figure 6 plots the forecasting results of the GM(1,1) model, TSI-GM(1,1) model, and LR model. Compared with the proposed model shown in Figure 5, the three models obviously exhibit poorer ability to follow the changes of SHPs' energy production. It can be also seen that, compared with the result in Yangbi County, both the GM(1,1) model and the LR model perform better than that in Yingjiang County, illustrating that these two models are more suitable for sequences with similar annual fluctuations. The TSI-GM(1,1) model can better identify the changing trend (upward/downward) of the next forecasting period than the GM(1,1) model and the LR model, but is insufficient in describing the quantity of change.

The performance evaluation criteria are given in Table 5. It can be seen that the proposed ISI-GM(1,1) model has the smallest RMSE, MAPE, MAE, and the biggest R^2 in both the study regions, exhibiting the best forecasting performance. The

absolute percentage error distributions of the different models in the two counties are given in Figures 7 and 8. The results indicate that the proposed model also shows more symmetrical error distribution. In addition, compared with the GM(1,1) model, the TSI-GM(1,1) model performs better in MAPE and R^2 , but worse in RMSE and MAE. It also indicated that TSI constructed from non-clustered local inflow data can only reflect the average fluctuation in multi-years rather than that of a given year. In some periods, by using the TSI, the seasonality of input data may not be weakened but even strengthened, which leads to larger forecasting errors.

As shown in Figures 5–8, for Yangbi County, the four models all perform worse in forecasting January 2015 to December 2015 compared with other periods, especially the LR model. The SHPs' monthly utilization hours are given in Figure 9 and the mean values are calculated from data from 2012 to 2014. It can be observed that, compared with 2012–2014, the flood period of Yangbi County in 2015 has a delay of one month. The SHPs annual utilization hours of Yangbi County in 2015 is 2,590 h, and the mean value is 3,539 h, i.e., dry year. As these models are built from historical data, their forecasting accuracies will decline when the next value deviates from its historical data. Therefore, the reason why all models perform worse in Yangbi County during January 2015 to December 2015 is that the observed values of 2015 deviate from the historical data. However, although the forecasting accuracy of the proposed ISI-GM(1,1) model in 2015 is also worse than other periods, it is the best among all the models. The performance evaluation criteria of the proposed model in forecasting 2015 are, RMSE = 51.18, MAPE = 13.69%, MAE = 32.33, and $R^2 = 0.938$, respectively, which is

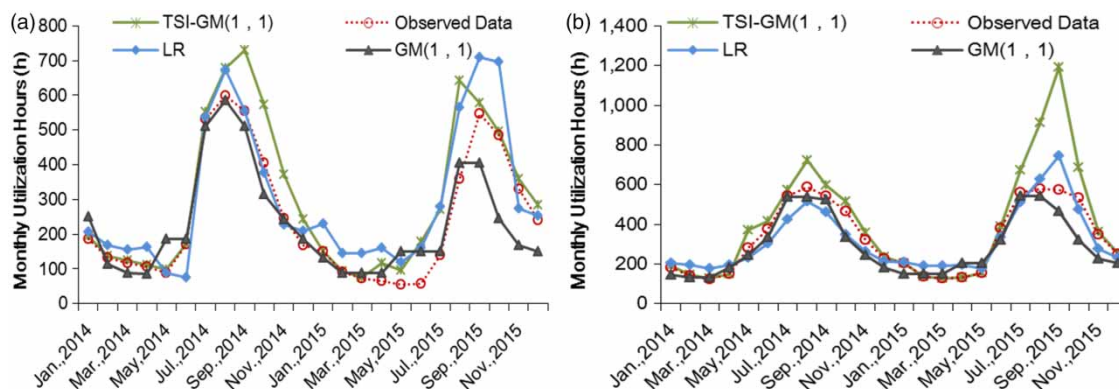


Figure 6 | Forecasting results of other compared models for Yangbi County and Yingjiang County during January 2014 to December 2015. (a) Yangbi County. (b) Yingjiang County.

Table 5 | Forecasting performances of different models for Yangbi and Yingjiang counties (RMSE, MAE: hour, MAPE: %)

Case study	ISI-GM(1,1) model			TSI-GM(1,1) model			GM(1,1) model			LR model			
	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE	R ²
Yangbi	38.71	9.93	23.38	93.25	33.53	59.86	81.18	35.43	56.8	91.96	47.56	69.86	0.865
Yingjiang	34.65	4.05	17.20	153.43	14.4	71.58	70.91	18.17	53.56	69.21	21.61	59.62	0.841

acceptable for practical engineering. In fitting precision checking, the C and P are 0.303 and 100.0%, showing a 'Good' grade. In contrast, for Yingjiang County, the energy production curves in 2014 and 2015 are similar and also approximate to the mean values, thus the forecasting performances of all models in 2015 have no obvious degradation. Above all, it can be concluded that by considering different inflow scenarios, the proposed ISI-GM(1,1) model has a better forecasting performance than the other three models and is more suitable for different inflow scenarios.

Discussion on cluster number

The forecasting performances with different cluster numbers used in the constructing process of ISI are discussed. The larger the cluster number is, the more groups that the reference LHP local inflow is divided into. In particular, when the cluster number is equal to 1, all observed values are treated as one subset and the ISI reflects the multi-year average. In this case, the ISI-GM(1,1) model is equivalent to the TSI-GM(1,1) model. The maximum cluster number in this study is equal to the local inflow data length, which means that each monthly observation is separately treated as a subset and the ISI is constructed for each year. The average forecasting errors for Yangbi and Yingjiang counties from January 2014 to December 2015 with different cluster numbers are shown in Figure 10. In both the study regions, the average forecast error is greatest when the cluster number is 1, drops to its minimum level when the cluster number is 3, then becomes larger and finally tends to be stable. Hence, in this paper, dividing the LHP local inflow data into three subsets is reasonable, which also agrees with the actual flow scenarios (i.e., wet, normal, and dry).

CONCLUSIONS

With more SHPs being integrated into the power grid, developing an effective forecasting model for SHPs' energy production is crucial for power systems operation and dispatching. However, most of the SHPs are located in remote areas and their historical information is absent. To overcome this problem, an original ISI-GM(1,1) model was proposed. The main contributions are summarized as follows:

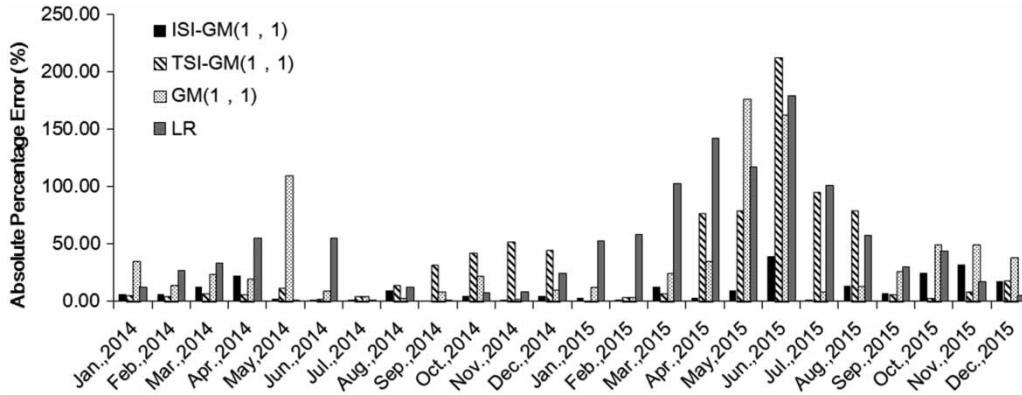


Figure 7 | Absolute percentage error distribution for Yangbi County during January 2014 to December 2015.

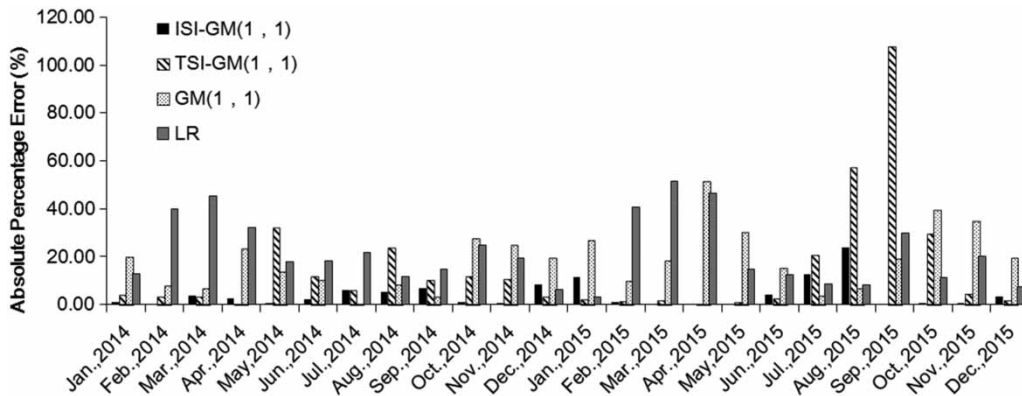


Figure 8 | Absolute percentage error distribution for Yingjiang County during January 2014 to December 2015.

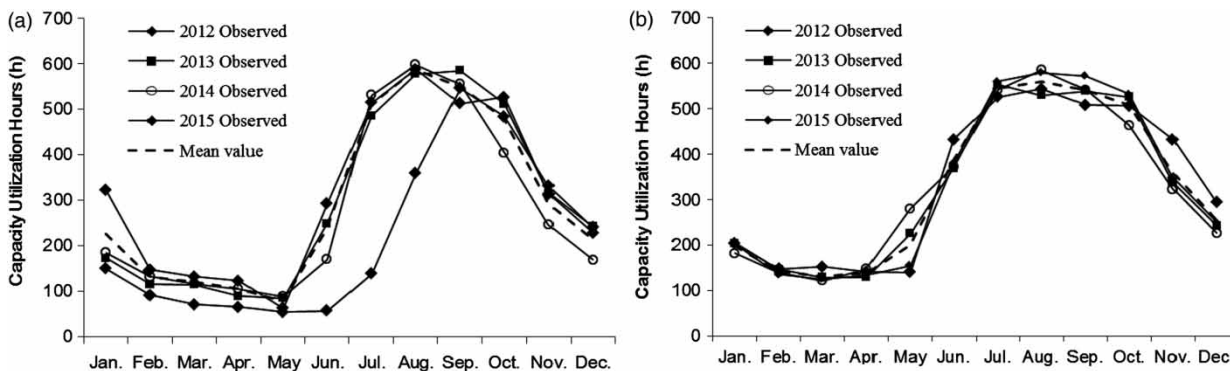


Figure 9 | The historical monthly utilization hours of Yangbi County and Yingjiang County. (a) Yangbi County. (b) Yingjiang County.

1. The correlation between LHP local inflow and SHPs' energy production was noted and analyzed, and then sufficient local inflow data from a reference LHP was employed in the energy production forecasting of SHPs.
2. An ISI was defined, and a hybrid model combining K-means clustering technique and RMA method was developed for calculating it. The simulation results show that the ISI can more reasonably reflect

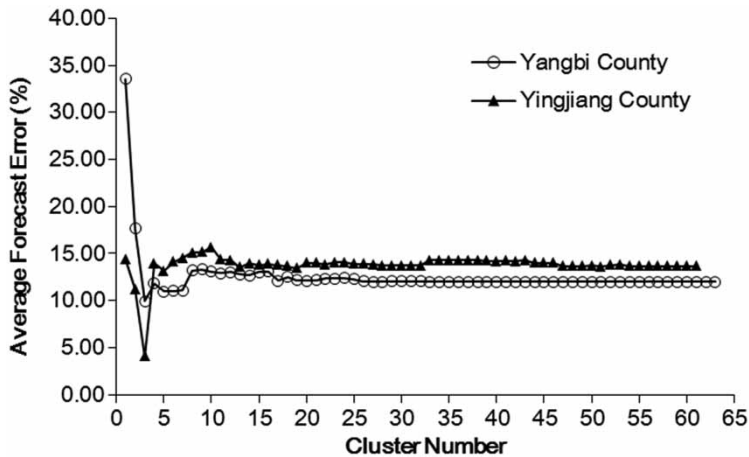


Figure 10 | Average forecasting errors of different cluster numbers used in constructing the process of ISI.

the seasonal variations of SHPs' energy production as compared with the traditional constant seasonal index.

- An ISI-GM(1,1) model was proposed by combining the GM(1,1) model with ISI to forecast monthly energy production for SHPs in ungauged basins, in which the ISI was introduced to enhance the forecasting accuracy of the GM(1,1) model with seasonal inputs. This paper offered a beneficial trial for GM in forecasting SHPs' energy production, and provided an alternative way for other seasonal time series prediction.
- The proposed ISI-GM(1,1) model was compared with the GM(1,1) model, the TSI-GM(1,1) model, and the LR model in forecasting the monthly energy production of SHPs in Yangbi and Yingjiang counties. The results show that the proposed model exhibited the best forecasting performance and was more suitable for different inflow scenarios, suggesting that the proposed ISI-GM(1,1) model is a feasible way to forecast monthly energy production of SHPs in ungauged basins.

It should be noted that this paper mainly focuses on a feasible way to forecast monthly energy production with a limited data set. The sensitivity and uncertainty of the proposed forecasting model and its input data were not considered and need to be further studied in the future, for example, the influence of the prediction error of reference LHP local inflow data and so on.

ACKNOWLEDGEMENTS

This work was supported by the Major Program of National Natural Science Foundation of China (No. 91547201), the National Basic Research Program of China (973 Program) (No. 2013CB035906), and the Major International Joint Research Project from the National Nature Science Foundation of China (No. 51210014).

REFERENCES

- Alvisi, S., Bernini, A. & Franchini, M. 2013 [A conceptual grey rainfall-runoff model for simulation with uncertainty](#). *Journal of Hydroinformatics* **15** (1), 1–20.
- Bai, Y., Chen, Z. Q., Xie, J. J. & Li, C. 2016 [Daily reservoir inflow forecasting using multiscale deep feature learning with hybrid models](#). *Journal of Hydrology* **532**, 193–206.
- Bakis, R. & Demirbas, A. 2004 [Sustainable development of small hydropower plants \(SHPs\)](#). *Energy Sources* **26** (12), 1105–1118.
- Bouckaert, R. R., Frank, E., Hall, M. A., Holmes, G., Pfahringer, B., Reutemann, P. & Witten, I. H. 2010 WEKA – experiences with a Java open-source project. *The Journal of Machine Learning Research* **11**, 2533–2541.
- Chau, K. W. & Wu, C. L. 2010 [A hybrid model coupled with singular spectrum analysis for daily rainfall prediction](#). *Journal of Hydroinformatics* **12** (4), 458–473.
- Cheng, C. T., Liu, B. X., Chau, K. W., Li, G. & Liao, S. L. 2015 [China's small hydropower and its dispatching management](#). *Renewable and Sustainable Energy Reviews* **42**, 43–55.
- Cheng, C. T., Luo, B., Miao, S. M. & Wu, X. Y. 2016 [Mid-term electricity market clearing price forecasting with sparse data: a case in newly-reformed Yunnan electricity market](#). *Energies* **9**, 804.

- Deng, J. L. 1982 The control problem of grey systems. *System & Control Letters* 1, 288–294.
- Deng, J. L. 1989 Introduction to grey system theory. *The Journal of Grey System* 1 (1), 1–24.
- Dudhani, S., Sinha, A. K. & Inamdar, S. S. 2006 Assessment of small hydropower potential using remote sensing data for sustainable development in India. *Energy Policy* 34 (17), 3195–3205.
- Estoperez, N. & Nagasaka, K. 2006 An artificial neural network based micro-hydropower generation scheduling: considering demand profile and environmental impact. *Clean Technologies and Environmental Policy* 8 (2), 123–130.
- Ferreira, J. H. I., Camacho, J. R., Malagoli, J. A. & Júnior, S. C. G. 2016 Assessment of the potential of small hydropower development in Brazil. *Renewable and Sustainable Energy Reviews* 56, 380–387.
- Golmar, G., Ramesh, R., Trevor, D., Pradeep, G. & Mari, V. 2017 Predicting the temporal variation of flow contributing areas using SWAT. *Journal of Hydrology* 547, 375–386.
- Hong, L. X., Zhou, N., Fridley, D. & Raczkowski, C. 2013 Assessment of China's renewable energy contribution during the 12th Five Year Plan. *Energy Policy* 62, 1533–1543.
- Kaunda, C. S. 2013 Energy situation, potential and application status of small-scale hydropower systems in Malawi. *Renewable and Sustainable Energy Reviews* 26, 1–19.
- Kim, S. E. & Seo, I. W. 2015 Artificial neural network ensemble modeling with exploratory factor analysis for streamflow forecasting. *Journal of Hydroinformatics* 17 (4), 614–639.
- Kong, Y. G., Wang, J., Kong, Z. G., Song, F. R., Liu, Z. Q. & Wei, G. M. 2015 Small hydropower in China: the survey and sustainable future. *Renewable and Sustainable Energy Reviews* 48, 425–433.
- Li, G., Liu, C. X., Liao, S. L. & Cheng, C. T. 2015a Applying a correlation analysis method to long-term forecasting of power production at small hydropower plants. *Water* 7 (9), 4806–4820.
- Li, G., Li, B. J., Yu, X. G. & Cheng, C. T. 2015b Echo state network with Bayesian regularization for forecasting short-term power production of small hydropower plants. *Energies* 8 (10), 12228–12241.
- Li, Z., Huang, G. H., Fan, Y. R. & Xu, J. L. 2015c Hydrologic risk analysis for nonstationary streamflow records under uncertainty. *Journal of Environmental Informatics* 26 (1), 41–51.
- Monteiro, C., Ramirez-Rosado, I. J. & Fernandez-Jimenez, L. A. 2013 Short-term forecasting model for electric power production of small-hydro power plants. *Renewable Energy* 50, 387–394.
- Monteiro, C., Ramirez-Rosado, I. J. & Fernandez-Jimenez, L. A. 2014 Short-term forecasting model for aggregated regional hydropower generation. *Energy Conversion and Management* 88, 231–238.
- Panwar, N. L., Kaushik, S. C. & Kothari, S. 2011 Role of renewable energy sources in environmental protection: a review. *Renewable and Sustainable Energy Reviews* 15 (3), 1513–1524.
- Purohit, P. 2008 Small hydro power projects under clean development mechanism in India: a preliminary assessment. *Energy Policy* 36 (6), 2000–2015.
- Renewable Energy Policy Network for the 21st Century (REN21) 2014 *Renewables 2014 Global Status Report*. Report, Renewable Energy Policy Network for the 21st Century, Paris, France.
- Srivastava, P. K., Han, D., Rico-Ramirez, M. A. & Islam, T. 2014 Sensitivity and uncertainty analysis of mesoscale model downscaled hydro-meteorological variables for discharge prediction. *Hydrological Processes* 28 (15), 4419–4432.
- Taele, B. M., Mokhutsoane, L. & Hapazari, I. 2012 An overview of small hydropower development in Lesotho: challenges and prospects. *Renewable Energy* 44, 448–452.
- Tan, P. N., Steinbach, M. & Kumar, V. 2014 *Introduction to Data Mining*, 1st edn. China Machine Press, Beijing, China, pp. 496–506.
- Taormina, R. & Chau, K. W. 2015 Neural network river forecasting with multi-objective fully informed particle swarm optimization. *Journal of Hydroinformatics* 17 (1), 99–113.
- Taylor, J. W. 2010 Triple seasonal methods for short-term electricity demand forecasting. *European Journal of Operational Research* 204 (1), 139–152.
- Tong, L. I., Saminathan, R. & Chang, C. W. 2016 Uncertainty assessment of non-normal emission estimates using non-parametric bootstrap confidence intervals. *Journal of Environmental Informatics* 28 (1), 61–70.
- Tongal, H. & Booiij, M. J. 2017 Quantification of parametric uncertainty of ANN models with GLUE method for different streamflow dynamics. *Stochastic Environmental Research and Risk Assessment* 31 (4), 993–1010.
- Tseng, F. M., Yu, H. C. & Tzeng, G. H. 2001 Applied hybrid grey model to forecast seasonal time series. *Technological Forecasting and Social Change* 67 (2), 291–302.
- Valipour, M., Banihabib, M. E. & Behbahani, S. M. R. 2013 Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. *Journal of Hydrology* 476, 433–441.
- Wang, W. C., Chau, K. W., Xu, D. M. & Chen, X. Y. 2015 Improving forecasting accuracy of annual runoff time series using ARIMA based on EEMD decomposition. *Water Resources Management* 29 (8), 2655–2675.
- Xie, N. M., Yuan, C. Q. & Yang, Y. J. 2015 Forecasting China's energy demand and self-sufficiency rate by grey forecasting model and Markov model. *International Journal of Electrical Power & Energy Systems* 66, 1–8.
- Yao, A. W. L., Chi, S. C. & Chen, J. H. 2003 An improved grey-based approach for electricity demand forecasting. *Electric Power Systems Research* 67 (3), 217–224.
- Yin, M. S. 2013 Fifteen years of grey system theory research: a historical review and bibliometric analysis. *Expert Systems with Applications* 40 (7), 2767–2775.
- Zhu, B. & Yuan, J. Q. 2015 Pressure transfer modeling for an urban water supply system based on Pearson correlation analysis. *Journal of Hydroinformatics* 17 (1), 90–98.