Environmental Research 184 (2020) 109262

Contents lists available at ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

Explore the relationship between fish community and environmental factors by machine learning techniques

Jia-Hao Hu[a], Wen-Ping Tsai[a,b,*], Su-Ting Cheng[c], Fi-John Chang[a,**]

[a] Department of Bioenvironmental Systems Engineering, National Taiwan University, No. 1, Roosevelt Rd., Taipei, 10617, Taiwan, ROC
[b] Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park, PA 16802-1408, USA
[c] School of Forestry and Resource Conservation, National Taiwan University, No. 1, Roosevelt Rd., Taipei, 10617, Taiwan, ROC

ARTICLE INFO

Keywords:
Sustainable environmental management
Water quality
Flow regime
Adaptive network-based fuzzy inference system (ANFIS)

ABSTRACT

In the face of multiple habitat alterations originating from both natural and anthropogenic factors, the fast-changing environments pose significant challenges for maintaining ecosystem integrity. Machine learning is a powerful tool for modeling complex non-linear systems through exploratory data analysis. This study aims at exploring a machine learning-based approach to relate environmental factors with fish community for achieving sustainable riverine ecosystem management. A large number of datasets upon a wide variety of eco-environmental variables including river flow, water quality, and species composition were collected at various monitoring stations along the Xindian River of Taiwan during 2005 and 2012. Then the complicated relationship and scientific essences of these heterogonous datasets are extracted using machine learning techniques to have a more holistic consideration in searching a guiding reference useful for maintaining river-ecosystem integrity. We evaluate and select critical environmental variables by the analysis of variance (ANOVA) and the Gamma test (GT), and then we apply the adaptive network-based fuzzy inference system (ANFIS) for an estimation of fish bio-diversity using the Shannon Index (SI). The results show that the correlation between model estimation and the biodiversity index is higher than 0.75. The GT results demonstrate that biochemical oxygen demand (BOD), water temperature, total phosphorus (TP), and nitrate-nitrogen (NO_3–N) are important variables for biodiversity modeling. The ANFIS results further indicate lower BOD, higher TP, and larger habitat (flow regimes) would generally provide a more suitable environment for the survival of fish species. The proposed methodology not only possesses a robust estimation capacity but also can explore the impacts of environmental variables on fish biodiversity. This study also demonstrates that machine learning is a promising avenue toward sustainable environmental management in river-ecosystem integrity.

1. Introduction

Concepts of conservation and restoration of natural ecosystems have gained an increasing interest in the last decades. As known, rapid urbanization, industrialization, and developments in catchment areas might have caused great impacts on, or even ruin, riverine ecosystems by producing high levels of nitrogen and phosphorus nutrients as well as inducing changes in water quality and flow patterns (Fashola et al., 2016; Förstner & Wittmann, 2012; Simmler et al., 2016). As such, exploring an effective ecosystem management plan requires careful consideration of the trends and changes in the environmental and biological conditions rooted in the long-term monitoring data.

Many studies have documented the crucial role of flow regime which is now viewed as an essential part of river ecosystem integrity (Acreman et al., 2014; Arthington et al., 2006 & 2010; Gillespie et al., 2015; Olden and Naiman, 2010; Papadaki et al., 2016; Tsai et al., 2016). In Taiwan, rivers possess high variations in their flow patterns from upstream to downstream, depending on the geographical condition as well as human activities. Moreover, precipitation patterns are highly uneven in space and time, with typhoons periodically occurring almost every year. These characteristics have been known to cause significant impacts on river ecosystems (Chang et al., 2015; Lee et al., 2016; Milliman et al., 2017). The Taiwan Eco-hydrologic Indicator System (TEIS) developed by Suen and Herricks (2006) connects hydrologic statistics with river organism requirements (fish species). The TEIS can be used as a tool to analyze the influences of hydrologic variation on communities of fish species (Chang et al., 2008 & 2013; Suen and Eheart, 2006). In addition to flow regime, water quality

* Corresponding author. Department of Bioenvironmental Systems Engineering, National Taiwan University, No. 1, Roosevelt Rd., Taipei, 10617, Taiwan, ROC.
** Corresponding author.
E-mail addresses: wptsai11@gmail.com (W.-P. Tsai), changfj@ntu.edu.tw (F.-J. Chang).

https://doi.org/10.1016/j.envres.2020.109262
Received 31 October 2019; Received in revised form 31 December 2019; Accepted 14 February 2020
Available online 17 February 2020
0013-9351/

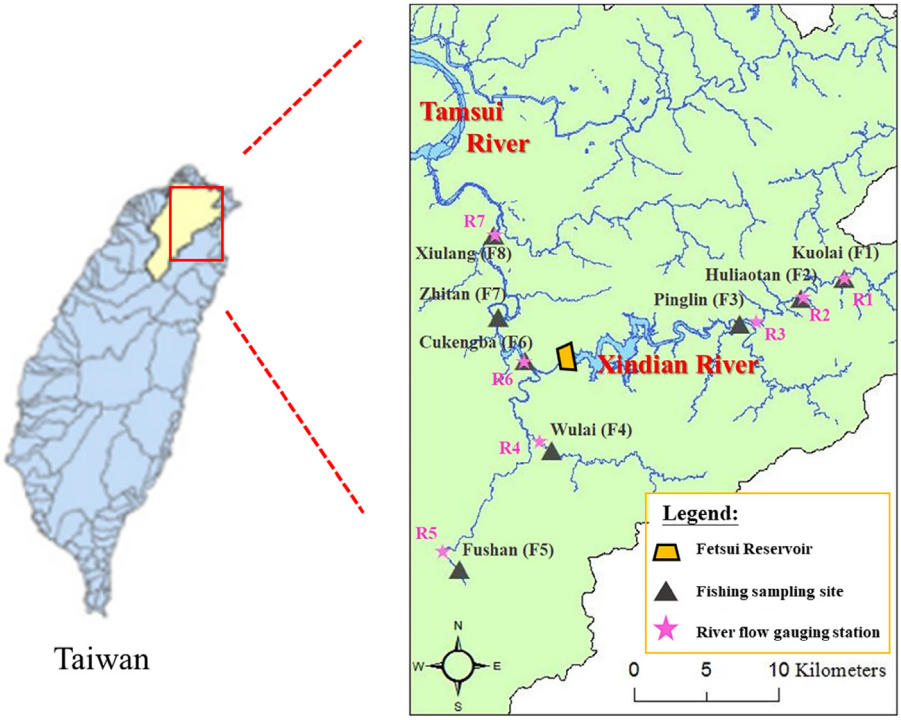


Fig. 1. Study area and locations of river flow gauge stations and fish sampling sites.

describing the physicochemical characteristics of a water body is undoubtedly a key factor affecting riverine ecosystems that water quality exacerbation due to industrial and agricultural development has been studied to assess its threat to certain fish species (Arthington et al., 2016; Hofmann et al., 2015; Liu et al., 2012; Tsai et al., 2017) or the cause of an alteration in fish communities (Chang et al., 2011; Cheng et al., 2018; Piperac et al., 2016; Segurado et al., 2016; Xu et al., 2016). Since flow regime and water quality are highly related to each other, any changes in either component may impose influences on the fish community to a certain extent (Cheng et al., 2016; Destouni et al., 2017; Marzin et al., 2013; Schinegger et al., 2012). For instance, Nilsson and Renöfält (2008) provided examples of how low flow conditions and changes in water quality could impact ecosystem processes and cause ecological problems. Whitehead et al. (2009) stated that increased water temperatures would affect chemical reaction kinetics and freshwater ecological status, and lower flows that reduce velocities and higher water residence times in rivers could induce a series of ecological responses.

However, actual applications in environmental management has been hampered by the lack of practical cases and methodologies. Most previous studies tended to assess the effects of a single factor on aquatic organisms and used statistical methods to obtain the correlation between environmental factors and fish communities (Kail et al., 2012; Kwon et al., 2012; Moerke and Lamberti, 2006), yet rare researches have considered the synergistic effects of multiple factors on fish biodiversity (Meng et al., 2009; Tsai et al., 2017; Zhao et al., 2018). Undoubtedly, there is an urgent need to conduct researches through novel data-driven techniques to comprehensively explore the complex interactions among various factors in an ecosystem and to model the effects of hydrological and water quality factors on the aquatic biota for suitably managing water resources and river ecosystems.

Machine learning techniques, such as artificial neural networks (ANNs), are known for their provision of reliable outputs through learning historical data. ANNs are predominant in extracting significant features from complex databases and are recognized for their outstanding abilities in modeling complex nonlinear systems. As a result, ANNs have been widely used for solving a wide range of fields with complex systems, such as hydrological (Chang et al., 2018; Chang and Tsai, 2016; Chen et al., 2018; Tsai et al., 2015; Uen et al., 2018) and eco-environmental fields (Barzegar et al., 2018; Chang et al., 2017; Dou and Yang, 2018; Forio et al., 2017; Halgamuge and Davis, 2019; Jones et al., 2017; Kaab et al., 2019; Nieto et al., 2013; Shi et al., 2018; Sannigrahi et al., 2019). Among them, the Adapted Network-Based Fuzzy Inference System (ANFIS) that combines fuzzy logic systems with a learning algorithm to construct the if-then rules and extract knowledge provides a promising alternative for feature extraction and prediction. The ANFIS has been satisfactorily applied to eco-hydrological and environmental fields (Blanes-Vidal et al., 2017; Hong et al., 2016; Nabavi-Pelesaraei et al., 2018; Woznicki et al., 2016; Yaseen et al., 2017; Zhou et al., 2019). However, many studies mentioned that overtraining could be one of the major weaknesses associated with modeling ANNs due to the improper selection of inputs and their data length for training (Chang et al., 2016; Remesan et al., 2009). The Gamma test (GT) introduced by Stefánsson et al. (1997) can successfully overcome these issues and sort input variables according to their importance and effects for training any smooth model before model construction. The GT has been used in various environmental studies (Chang et al., 2016; Goyal et al., 2014; Noori et al., 2015; Remesan et al., 2008; Tian et al., 2016), which has implied the GT could help minimize the guesswork in modeling ANNs by identifying critical input variables.

To achieve sustainable environmental management, it is necessary to consider the uniqueness of the physical habitat environment and the relationship to their inhabitant biota on a local scale. In this study, we intend to develop a sophisticated methodology capable of extracting knowledge from limited heterogeneous datasets to form if-then rules for judging the associated fish biodiversity, with an aim to build an ANFIS model for estimating fish biodiversity based on environmental variables (flow regime and water quality). Specific objectives are: (1) identify key environmental variables effectively for modeling the complex eco-hydrological system by the GT; (2) construct a reliable ANFIS model for estimating fish biodiversity; (3) explore the relationship between environmental variables and fish biodiversity based on the if-then rules of the ANFIS; and (4) provide a guiding reference for decision-makers to

maintain the integrity of a river ecosystem.

2. Study area and data collection

2.1. Study area

The Xindian River is one of the three major tributaries of the Tamsui River in northern Taiwan (Fig. 1). With a length of 81 km and a large catchment area of 921 km^2, the Xindian River flows through the New Taipei City and the Taipei City. In this watershed, the landform has a considerable variation in elevation and the climatic feature is that it is usually wet and rainy. The Beishi River is one of the main tributaries in the upstream, where tea orchards and tea leisure farms are the main land-use types. The Feitsui Reservoir is situated in the middle of the Xindian River to meet the water demand of over 4 million residents in the vast Taipei region.

2.2. Data collection

To explore the complicated relationships among environmental factors and fish bio-diversity, we collected data of river flow (daily), water quality (monthly), and fish sampling across the Xindian River basin during 2005 and 2012. These data were monitored by different governmental agencies. To describe the river status and compared it with other environmental variables under the same time scale, we converted the flow data into monthly flow regimes by a total of 9 Taiwan Eco-Hydrologic Indicator System (TEIS; Suen and Herricks, 2006; Chang et al., 2013; Tsai et al., 2015 & 2016) variables (Table 1).

As hydro-chemo-biological factors are known to control fish species presence and abundance as well as shape the community structure along with the unidirectional river networks (Cheng et al., 2018), we gathered a total of 134 fish survey samples at eight sampling sites (Table 2). Fish surveys were conducted by either electroshock or creel surveys at eight sampling sites. Meanwhile, temperature (Temp), hydrogen ions concentration (pH), electric conductivity (EC), turbidity, dissolved oxygen (DO), suspended solids (SS), biochemical oxygen demand (BOD), ammonia nitrogen (NH_3–N), nitrate-nitrogen (NO_3–N) and total phosphorus (TP) were also collected (Table 3) at the same periods of the fish surveys.

3. Methods

To assess the complex relationships among river flow, water quality, and fish biodiversity for providing guiding references of species conservation, we proposed a sophisticated three-phase data-mining methodology comprising the analysis of variance (ANOVA), the Gamma test (GT), and the Adapted Network-Based Fuzzy Inference System (ANFIS) (Fig. 2). We first converted streamflow and fish sampling data into flow regime and fish biodiversity index, respectively (Section 3.1). Following that, the ANOVA (Section 3.2) was used to capture the environmental divergence of fish sampling sites, and the GT (Section 3.3) was performed to identify the key variables as the model inputs for reducing model complexity and increasing model stability and reliability. Then the ANFIS (Section 3.4) model was trained and validated to increase model accuracy and scalability. At last, we explored the relationships among selected variables using the if-then rules revealed by the ANFIS model (Fig. 2).

Table 1
Flow regime in a monthly scale.

Flow regime (monthly scale)	Variables
TEIS m1	Mean of all positive differences between consecutive values in the month corresponding to fish sampling.
TEIS m2	Mean of all negative differences between consecutive values in the month corresponding to fish sampling.
TEIS m3	Mean streamflow of the respective month.
TEIS m4	1-day maximum streamflow of the respective month.
TEIS m5	1-day minimum streamflow of the respective month.
TEIS m6	3-day maximum streamflow of the respective month.
TEIS m7	3-day minimum streamflow of the respective month.
TEIS m8	10-day maximum streamflow of the respective month.
TEIS m9	10-day minimum streamflow of the respective month.

Table 2
Fish sample sites and corresponding survey years.

Sampling site	Code	Survey year	Number of data
Kuolai	F1	2005–2009, 2011–2012	24
Huliaotan	F2	2011–2012	8
Pinglin	F3	2005–2009, 2011–2012	24
Wulai	F4	2005–2009, 2011–2012	24
Fushan	F5	2005–2009, 2011–2012	24
Cukengba	F6	2005–2009, 2011–2012	24
Zhitan	F7	2005	2
Xiulang	F8	2004, 2005	4

Table 3
Basic statistics of all variables in the Xindian River basin.

	Minimum	Maximum	Median	Average	Standard Deviation
Temp (°C)	11.20	32.00	22.65	22.90	3.99
pH	6.30	29.80	7.66	7.86	1.99
EC (μS/cm)	8.40	291.00	83.00	90.26	29.26
Turbidity (NTU)	0.00	230.00	1.60	10.55	30.64
DO (ppm)	3.30	11.20	8.48	8.30	1.28
SS (ppm)	0.15	300.00	1.60	14.75	43.38
BOD (ppm)	0.01	3.40	0.80	0.88	0.56
NH_3–N (ppm)	0.00	3.13	0.05	0.09	0.28
NO_3–N (ppm)	0.00	4.59	0.38	0.41	0.43
TP (ppm)	0.00	0.52	0.03	0.04	0.07
TEIS m1	0.00	326.99	9.94	25.27	43.34
TEIS m2	0.02	122.82	5.64	13.23	19.97
TEIS m3	0.27	198.60	23.37	36.01	34.47
TEIS m4	1.47	2048.61	83.54	191.21	299.26
TEIS m5	0.00	78.11	6.18	12.40	14.95
TEIS m6	1.04	984.56	64.17	114.32	149.24
TEIS m7	0.00	78.41	8.62	14.85	16.48
TEIS m8	0.55	364.43	36.70	60.39	63.27
TEIS m9	0.00	103.33	11.89	19.46	20.63
Shannon Index	0.00	2.73	1.48	1.48	0.46

3.1. Fish diversity index – Shannon Index (SI)

The SI was proposed by Shannon (1948) as a common indicator to assess ecological diversity (Keylock, 2005; Spellerberg and Fedor, 2003). It has been widely used to estimate the mutual proportion of species in a community under the assumption that an individual is randomly investigated from an infinite community, and can be calculated as:

$$SI = -\sum_{i=1}^{S} p_i \ln p_i \quad (1)$$

where S is the number of species of the community, and p_i is the proportion of the ith species to the total number of individuals.

3.2. Statistical technique - analysis of variance (ANOVA)

The ANOVA is a widely used procedure to partition observed variances into different explanatory variables, which helps to determine the relevant statistical significance (Lindman, 1974). To compare the total deviation of the variables analyzed in the ANOVA, the F-test is conducted using the following equation.

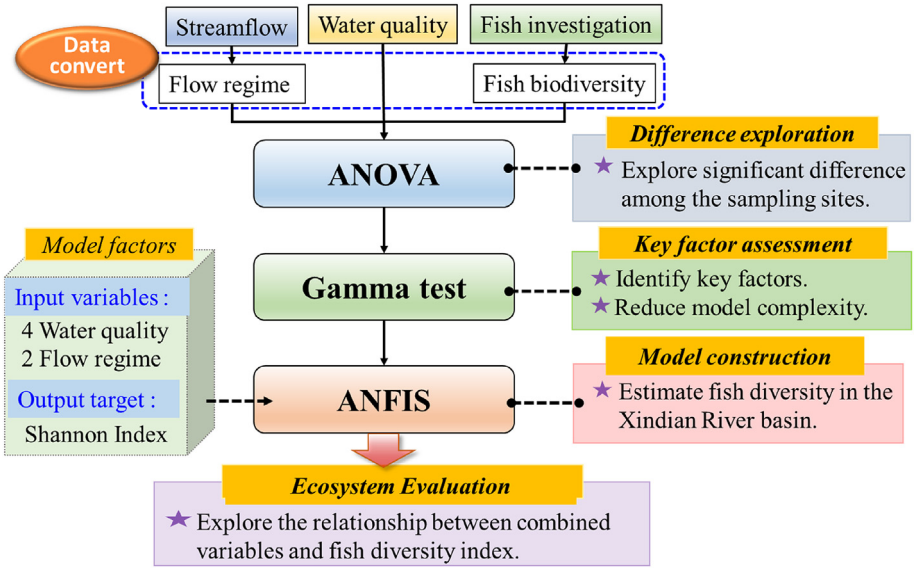


Fig. 2. Study framework.

$$F=\frac{\sum_{i=1}^{I} n_i(\bar{Y}_i-\bar{Y}_{total})^2}{\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\bar{Y}_i)^2} \quad (2)$$

where I is the total number of groups in the data set, J is the total number of observational data, Y_{ij} is the jth observational data of the ith group, and $\bar{Y}_{total}$ is the average of the whole observational data. Then, the probability (p-value) of a value of F is calculated to check if it is greater than or equal to the observed value. The null hypothesis is rejected if this probability is less than or equal to a commonly used significance level (α) of 0.05.

3.3. Gamma test (GT)

The GT proposed by Stefánsson et al. (1997) has been utilized to select essential inputs by estimating the noise level presented in a data set and to identify the best input variables without extensive model development for each potential input combination. This technique is used to select effective variables for modeling the highly non-linear eco-environmental problem in this study. When the effective variables are selected to form an input combination, they can reduce model complexity and promote model reliability (Moghaddamnia et al., 2008; Noori et al., 2011). To employ the GT, an observational data set of input-output patterns can be described as:

$$\{(x_i, y_i): 1 \le i \le M\} \quad (3)$$

where the inputs $x_i \in R^m$ are m dimensional vectors with M length confined to a certain closed bounded set $C \subset R^m$, and the corresponding outputs $y \in R$ are scalars.

The relationship of input-output patterns can be shown as:

$$y = f(x) + r \quad (4)$$

where f is a smooth function representing the system, and r is the noise of a random variable.

Based on the Euclidean distance of the k^{th} nearest neighbors $X_{N(i,k)}$ for each vector X_i, the Delta function is defined by the following equation.

$$\delta_M(k) = \frac{1}{M}\sum_{i=1}^{M} |X_{N(i,k)} - X_i|^2 \quad (1 \le k \le p) \quad (5)$$

where p is the number of neighboring points.

The Gamma function of the output values is given as:

$$\gamma_M(k) = \frac{1}{2M}\sum_{i=1}^{M} |y_{N(i,k)} - y_i|^2 \quad (1 \le k \le p) \quad (6)$$

where $y_{N(i,k)}$ are the y-values corresponding to the inputs X_i, which are the k^{th} nearest neighbors in the input domain.

A regression line is constructed for the p points $(\delta_M(k), \gamma_M(k))$, shown as follows.

$$\gamma = A\delta + \Gamma \quad (7)$$

where A is the gradient. Γ, known as the Gamma statistic, is the intercept of the regression line of $\gamma_M(k)$ versus gh $\delta_M(k)$, which shows the noise estimate for each subset of input variables. The bigger the Γ value is, the higher the complexity of the model is. In other words, if the Γ value is big, the corresponding input combination can be regarded as the worse combination; and if the Γ value is the closest to zero, the corresponding input combination can be regarded as the best combination.

3.4. Adaptive network-based fuzzy inference system (ANFIS)

ANNs can usually achieve high estimation accuracies but have a drawback of lacking explainability, which significantly limits their applicability (Mount et al., 2016). The ANFIS introduced by Jang (1993) considers the fuzzy inference system as a core fundamental and combines it with the ANN for providing qualitative description and reasoning processes of human-knowledge (Chang et al., 2005). This study proposes a methodology that is capable of extracting rule-based knowledge according to the input-output relation of a trained ANFIS model for assessing the impacts of environmental variables on fish biodiversity. The framework of the ANFIS (Fig. 3) includes five layers: an input layer, a rule layer, an average layer, a consequent layer, and an output layer. In this study, the ANFIS applies the Takagi-Sugeno fuzzy model (TSK fuzzy model) to configuring the if-then rules of the fuzzy inference system (the rule layer) and uses similar membership functions in the same layer to build the principal structure. Taking two inputs, x_1 and x_2, in a fuzzy inference system and an output, y as an example, in the first-order TSK fuzzy model, a simple rule set with two fuzzy if-then rules can be described as:

Rule 1: If x_1 is A_1 and x_2 is B_1 then $y = p_1 * x_1 + q_1 * x_2 + r_1$ (8)

Rule 2:If x_1 is A_2 and x_2 is B_2 then $y = p_2 * x_1 + q_2 * x_2 + r_2$

where p, q, and r are linear parameters in the then-part of the first-order

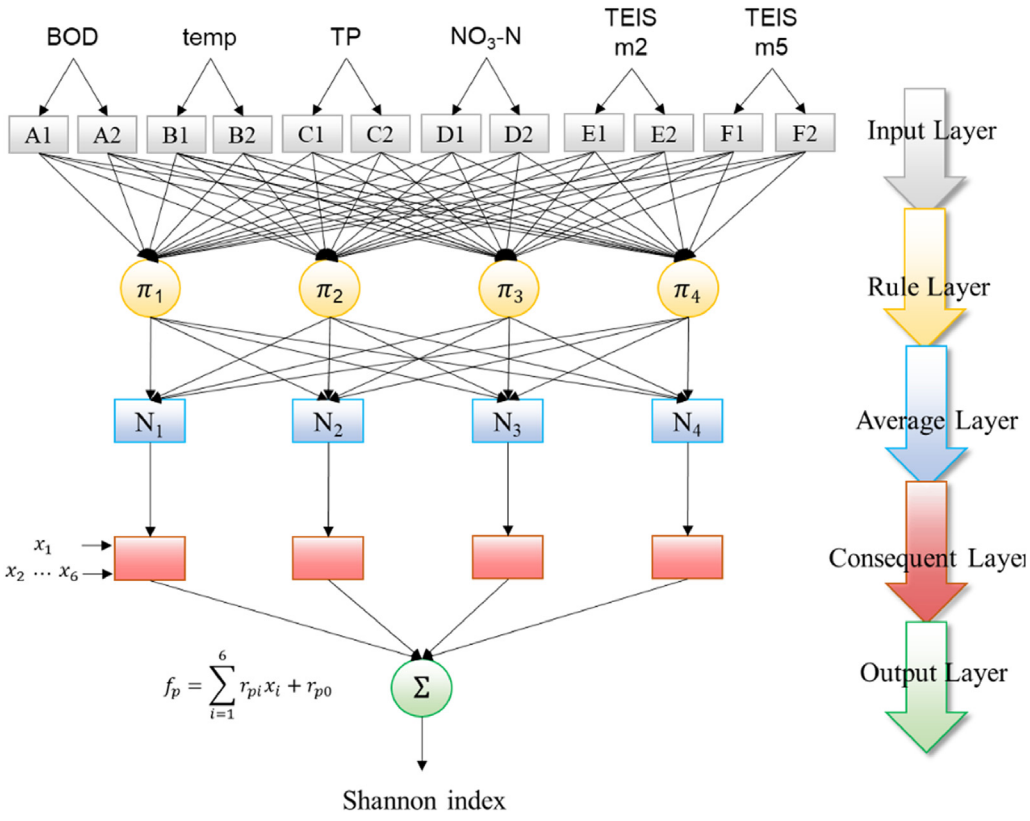


Fig. 3. Framework of the ANFIS.

TSK fuzzy model. By connecting the feedforward networks and the supervised learning algorithm, the ANFIS can adjust all the parameters properly to make itself have the ability of self-learning and self-organizing.

In this study, we focused on what selected variables cause the difference between the ANFIS rule-base. As we mentioned earlier, species diversity can be considered as a kind of ecosystem integrity. Two main environmental factors, i.e., flow regime and water quantity, in the river ecosystem of this study were selected and qualified as the ANFIS inputs. We assessed the species diversity condition using the ANFIS and discussed the connection between input variables and species diversity.

4. Results and discussion

This study proposes a machine learning methodology for exploring the complex impacts of water quality and flow regime on fish communities to comprehend the situations of the eco-hydrological system in a river basin and to resolve the complex and interrelated conservation and management issues.

4.1. Results of the ANOVA

In this study, we collected a total of 134 datasets, including water quality, flow regime, and fish survey at 8 fish sampling sites (Table 2). However, three fish sampling sites (Huliaotan (F2), Zhitan (F7), and Xiulang (F8)) with limited datasets did not have the same number of datasets as the other five sampling sites. Thus, we removed F2, F7, and F8 from the ANOVA analysis. The remaining five sampling sites (Kuolai (F1), Pinglin (F3), Wulai (F4), Fushan(F5), and Cukengba (F6), with a total of 120 data sets (24 datasets from each sampling site), were used to explore the differences of water quality factors and flow regime among sampling sites using the ANOVA. The results were summarized in Table 4.

It indicates that three water quality factors of DO, BOD, and NH_3–N are not significantly different in sampling sites. Moreover, the high DO (8.1–8.5 ppm) and low BOD (0.8–0.9 ppm) at all the five stations suggest good water quality. Thus, the fish species diversity (1.31–1.87) is high overall. In the aspect of geographical difference, some water quality variables, such as EC and SS, do make a significant difference between different tributaries. The temperature variable also presents a significant difference between F3 and F5, where the elevation of F5 is significantly higher than that of F3. According to the results of SS, F5 performs significantly different from F1, F3, and F4. We notice that F5 has the highest SS concentration among sampling sites, which can be because the frequent landslides in the surrounding area of F5 slump eroded soil into the river channel when large rainfall events occur. Concerning the impacts of human activities on sampling sites, the NO_3–N variable, whose main source is fertilizer, at F3 has a significant difference with the other sites. The main reason might be that the area near F3 locates a large tea plantation, where fertilizers are needed to provide tea trees with nutrition.

Most flow regime variables differ significantly between different sampling sites (Table 4), except for TEIS m1 and TEIS m4. The SI values are higher at F3 and F6 but lower at F4 and F5, which implies a significant difference between sites. According to the ANOVA results, the performance of F3 is significantly different from those of F1, F4, and F5, while the performance of F6 is significantly different from those of F4, and F5. The main reason that caused lower SI values at F1, F4, and F5 should be the higher elevations of these sites. This result provides an extra evidence to support the previous studies (Allouche et al., 2012; Hortal et al., 2013; Lawton et al., 1987) that claims a site located at a higher-elevation usually contains fewer fish species due to the area available per habitat decreases (shallow surface water, and narrow river channels).

The ANOVA results of water quality variables not only reflect the spatial (geographical) differences but also imply the effects of human activities on sampling sites. Based on the analytical results, we notice that (1) good water quality (high DO and low BOD) would commonly

Table 4
ANOVA results.

Sub-basin		Beishi River		Neidong River	Nanshi River	Downstream
Site		F1	F3	F4	F5	F6
Variables	Elevation (m)	250	180	200	390	100
Temp (°C)	Mean	23.3	24.4	22.8	20.8	22.2
	S.D.[a]	–	(F5)[b]	–	(F3)	–
pH	Mean	7.8	7.8	7.9	7.6	7.36
	S.D.	–	–	(F6)	–	(F4)
EC (μS/cm)	Mean	67.8	87.5	80.6	112.2	100
	S.D.	(F3, F5, F6)	(F1, F5)	(F5, F6)	(F1, F3, F4)	(F1, F4)
Turbidity	Mean	1.4	4.2	1.7	28.6	22.1
	S.D.	(F5)	–	(F5)	(F1, F4)	–
DO (ppm)	Mean	8.1	8.32	8.3	8.3	8.5
	S.D.	–	–	–	–	–
SS (ppm)	Mean	0.98	3.4	1.42	33.1	19.5
	S.D.	(F5)	(F5)	(F5)	(F1, F3, F4)	–
BOD (ppm)	Mean	0.8	0.9	0.8	0.8	0.8
	S.D.	–	–	–	–	–
NH_3–N (ppm)	Mean	0.05	0.07	0.06	0.07	0.07
	S.D.	–	–	–	–	–
NO_3–N (ppm)	Mean	0.28	0.69	0.41	0.32	0.44
	S.D.	(F3)	(F1, F5)	–	(F3)	–
TP (ppm)	Mean	0.02	0.03	0.03	0.06	0.07
	S.D.	(F6)	–	–	–	(F1)
TEIS m1	Mean	9.26	15.31	33.03	32.87	32.52
	S.D.	–	–	–	–	–
TEIS m2	Mean	4.78	9.25	16.14	8.95	21.48
	S.D.	(F6)	–	–	–	(F1)
TEIS m3	Mean	11.93	16.61	52.97	35.85	61.22
	S.D.	(F4, F5, F6)	(F4, F6)	(F1, F3)	(F1)	(F1, F3)
TEIS m4	Mean	61.14	117.11	250.69	170.29	276.37
	S.D.	–	–	–	–	–
TEIS m5	Mean	3.32	3.2	19.52	14.8	22.99
	S.D.	(F4, F5, F6)	(F4, F5, F6)	(F1, F3)	(F1, F3)	(F1, F3)
TEIS m6	Mean	40.52	66.16	147.80	105.69	173.4
	S.D.	(F4, F6)	–	(F1)	–	(F1)
TEIS m7	Mean	3.72	3.93	24.63	16.54	26.78
	S.D.	(F4, F5, F6)	(F4, F5, F6)	(F1, F3)	(F1, F3)	(F1, F3)
TEIS m8	Mean	21.54	31.13	81.28	60.02	98.87
	S.D.	(F4, F6)	(F4, F6)	(F1, F3)	–	(F1, F3)
TEIS m9	Mean	5.3	5.51	32.21	21.2	34.98
	S.D.	(F4, F5, F6)	(F4, F5, F6)	(F1, F3)	(F1, F3)	(F1, F3)
Shannon Index	Mean	1.51	1.87	1.31	1.31	1.66
	S.D.	(F3)	(F1, F4, F5)	(F3, F6)	(F3, F6)	(F4, F5)

[a] Significant Difference.
[b] Each sampling site in a bracket shows a significant difference with the sampling site of the column that it is positioned.

lead to higher fish species diversity, (2) the highest SS concentration at F5 is caused mainly by the frequent landslides in the surrounding area, and (3) the NO_3–N variable at F3 has a significant difference with those of the other sites because F3 locates a large tea plantation, where fertilizers are needed.

4.2. Input combination selection - results of the GT

Despite a great number of studies on modeling ANNs, there are still some unresolved issues, such as the identification of input factors that are more relevant to estimation/prediction. Overtraining is considered as another severe weakness encountered during ANN model construction, where excellent results are created by the training data but poor results are produced by the unseen test data because of the improper selection of inputs and/or the data length for training. The GT can adequately overcome these issues and is used in this study to determine the non-trivial key input items for modeling the ANFIS to achieve reliable fish bio-diversity estimation. Initially, 10 water quality variables and 9 water quantity (flow regime) variables are considered as input items. However, 19 input variables are too complicated to implement the ANFIS model, especially based only on 134 datasets. It is necessary to adopt the GT to identify proper input variables. Thus, there are a large number of possible combinations of input variables to determine the key factors for the ANFIS model.

Before conducting the GT, heterogeneous data of water quality and water quantity were normalized to [-1, 1] in avoidance of the bias raised from the difference in scales. The GT produced a total of 1023 (2^{10}–1) Γ values associated with 10 water quality (i.e., 1023 combinations of variables) and a total of 511 (2^9–1) Γ values associated with 9 water quantity (511 combinations of variables). In other words, each Γ value engages a combination of variables. Then, Γ values are sorted in ascending order, in which the combinations corresponding to Γ values smaller than the 10th percentile are classified as the best combination group, whereas the combinations corresponding to Γ values bigger than the 90th percentile are classified as the worst combination group. Finally, the ratio of the occurrence frequency of each variable in the best combination group to that of the worst combination group is calculated to determine the most suitable input combination for the ANFIS. Fig. 4 shows the results of the GT, where a blue bar denotes the occurrence frequency of a variable in the best combination group, a red bar denotes the occurrence frequency of a variable in the worst combination group, and a black point presents a ratio defined above for a variable. According to Fig. 4, we chose the elbow (significant turning point) of the black curve as the threshold to select input variables. It indicates that BOD, temp, TP, and NO_3–N (water quality) as well as TEIS m2 and TEIS m5 (flow regime) form the optimal input

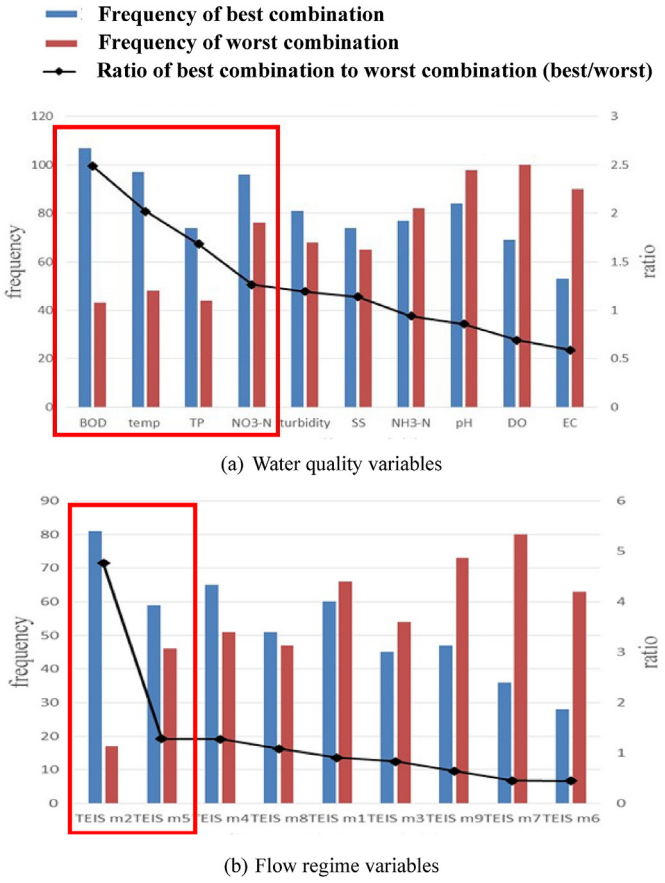


Fig. 4. Determination of ANFIS input items by the GT. (a) Water quality variables. (b) Flow regime variables.

combination of the ANFIS. Fig. 4 also demonstrates that the GT is a tool suitable for identifying the relationship between input combinations and output item.

Concerning water quality, although each of the pairs [BOD, DO] and [NO_3–N, NH_3–N] shows similar ANOVA results (no significant difference in spatial distribution), the GT suggests BOD and NO_3–N are more important factors affecting fish biodiversity, rather than DO and NH_3–N. This expresses that the GT can discover valuable information that might not be detected by traditional statistical methods. Moreover, the GT also reduces the noise present during ANN modeling and mitigates the problem of overfitting.

4.3. Fish biodiversity estimation – results of the ANFIS

This study estimates fish diversity indexes (Shannon Index, SI) using the ANFIS based on key input factors determined by the GT. For building a robust model under the condition of fish data scarcity, we need to train the model based on the training datasets and test it on some unseen test datasets. Cross-validation is a resampling procedure commonly used in machine learning modeling to estimate the skill of a constructed model on unseen data when only limited data are available (Varma and Simon, 2006). This popular method is easy to understand and generally leads to a less biased estimate of the model skill than other methods, such as a simple split of datasets into training and test ones. Because we only had limited datasets obtained from eight monitoring sites, we used the cross-validation procedure to train and validate the ANFIS models. We randomly shuffled (re-sample) the complete datasets (134 datasets) into training (120) and testing sets (14), and the shuffling procedure was executed 1000 times to achieve the desired accuracy and reliability.

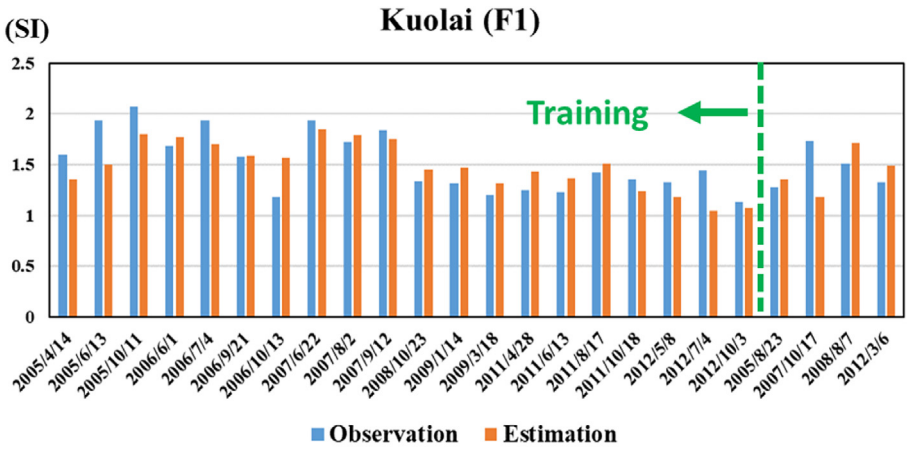
Table 5
Model performance concerning the estimation of the Shannon Index.

	RMSE	MAE	CC	CE
Training	0.30	0.23	0.78	0.60
Testing	0.30	0.24	0.75	0.56

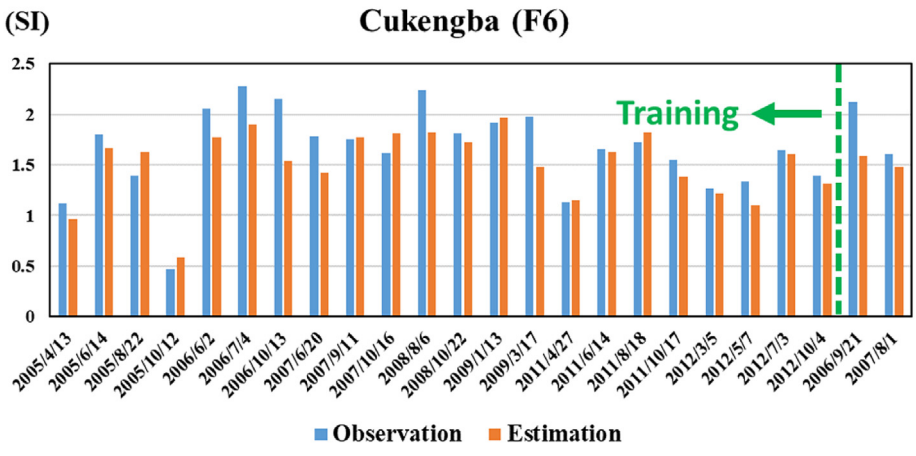
Four common criteria, i.e., root mean square error (RMSE), mean absolute error (MAE), coefficient of efficiency (CE), and coefficient of correlation (CC) are applied to evaluating model performance. Table 5 presents the performance of the estimated SI produced by the ANFIS in terms of the four evaluation criteria. We note that even the fish biodiversity (SI) is significantly different from site to site and results in a large overall standard deviation (0.46, Table 3), the fish biodiversity still can be estimated well through the machine learning technique based on two heterogeneous datasets (flow regime and water quality). The results reveal that the ANFIS combined with the GT can provide robust and stable performance of estimation, with the CC reaches about 0.78 and 0.75 in the training and testing stages, respectively, and the MAE stays only at 0.23 and 0.24 (about half of standard deviation) in the training stage and testing stage, respectively. Fig. 5 shows the estimation results of the ANFIS at F1 (upstream reach, Fig. 5(a)) and F6 (downstream reach, Fig. 5(b)). We can find that the ANFIS model has good performances in estimating fish biodiversity in both up- and downstream reaches. The results demonstrate that the ANFIS combined with the GT can provide accurate estimations of fish diversity, with high CC (exceeding 0.75) and low RMSE/MAE (half of the standard deviation) over different spatial and temporal distributions.

4.4. Linkage of environmental and water quality variables with fish biodiversity – membership functions

Despite the input combination determined by the GT, the principal structure of the ANFIS model, i.e., if-then rules of fuzzy inference system, can offer further interpretations between input items and the output. After verifying the reliability of the ANFIS model in fish biodiversity estimation, this study probes into the cognitive construction of the fuzzy sets of the individual selected input variables by plotting membership functions (MFs). The membership functions corresponding to an input variable visualize the auxiliary explanation to address the impact of this variable on the model target. Fig. 6 shows the MFs of each selected input variable and the model target (SI). We can find that the MFs of an input can be classified into different distributions, while most of these input variables do not show significant differences in the ANOVA. The results indicate the powerful ability of the ANFIS combined with the GT in feature extraction. For instance, the subtle differences between variables cannot be detected by the ANOVA but can be interpreted by the patterns/distributions of the MFs produced from the ANFIS. From the results of individual single factor shown in Fig. 6(a)-6(f), it appears that the patterns of the MFs corresponding to Temperature, BOD, and NO_3–N are similar to those of the model target (SI), especially for the amplitude of MF1 (wider than those of the other three MFs). These distributions reflect that the fish species distribution in the study area highly depends on the water temperature, BOD, and NO_3–N (consistent with the previous study, Chen, 2009 & 2011). Furthermore, the highest MF values of BOD (Fig. 6(b)) and TEIS m2 (Fig. 6(e)) are engaged with Rule 3 (MF3), which would result in the lowest MF value of fish biodiversity (Fig. 6(g)). In other words, higher BOD implies the water body might be polluted whereas higher TEIS m2 means lower streamflow in the river (habitat shrinks). Based on Rule 3 (MF3), the lowest fish biodiversity reports that it is difficult for most of the fish species to survive under such environmental conditions. Besides, Fig. 6 also points out that the highest MF values of TP (Fig. 6(d)) and TEIS m5 (Fig. 6(f)) are contributed by Rule 2 (MF2), which might lead to higher fish biodiversity (Fig. 6(g)). The implicit reason is that TP



(a) Kuolai (F1)



(b) Cukenba (F6)

Fig. 5. Estimation results of the ANFIS.

is the main nutrition source for plankton and some fish species, while plankton is the main nutrition source for some fish species. Thus, the conditions of Rule 2 (MF2) here present the advantages of a higher nutrition source of TP, while higher TEIS m5 means larger habitats would be more suitable for fish species to survive.

This study demonstrates that the ANFIS has high flexibility to build a rule base grounded on historical records and can configure a powerful machine learning framework to deal with heterogeneous datasets. The visualization of membership functions (MFs) is one of the essences inhered in the ANFIS. The results show that the ANFIS can effectively extract the relationship between environmental variables and fish communities and offer a better understanding of how environmental variables affecting fish communities in light of MFs. The MFs not only can properly reflect the linkages between input variables and the model target but also can identify the situation of fish biodiversity in a complex environmental background.

5. Conclusions

Machine learning techniques are efficient and scalable tools for data analysis and pattern recognition and have been widely used in the environmental field. The practices of machine learning techniques, however, commonly face awkward situations engaging a large number of potential inputs but limited datasets when trying to construct reliable and interpretable models. This study proposed a hybrid approach that integrates a machine learning model and the Gamma test not only to extract valuable and meaningful information from the selected environmental variables but to explore the impacts of these environmental variables on the river fish biodiversity through the underlying connections of the machine learning framework. Besides, the cross-validation was carried out to improve the reliability and robustness of the constructed model on unseen data under the condition of fish data scarcity. The heterogeneous monitoring datasets of river flow, water quality, and fish sampling collected over the Xindian River basin in the northern Taiwan during the period from 2005 to 2012 formed a case study.

The statistical significance of environmental variables revealed from the ANOVA indicates the environmental variables (water quality and flow regime factors) and the ecosystem index (fish bio-diversity) exhibit inevitable divergence on account of geographical differences as well as human activities, which could affect the reliability and accuracy of estimation models.The importance of four water quality factors (BOD, Temp, TP, and NO_3–N) and two flow regime factors (the decreasing rate of streamflow and 1-day minimum streamflow) to fish biodiversity in the study area was identified by the GT. A more suitable environment for the survival of fish species can generally be created under the condition of lower BOD, higher TP, and larger habitat (flow regimes), suggested by the ANFIS model.

In summary, the proposed methodology can effectively identify important environmental variables for modeling fish biodiversity,

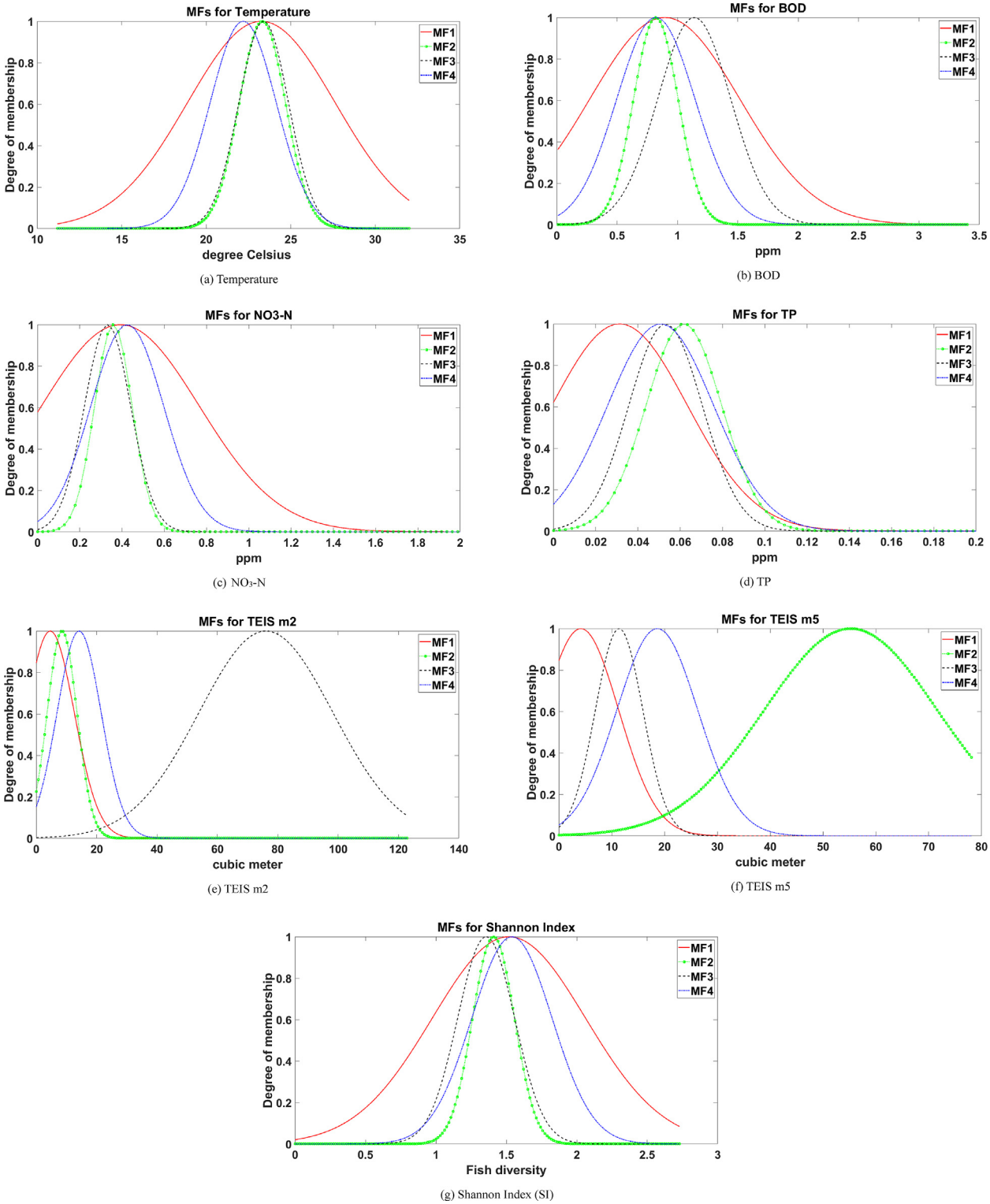


Fig. 6. Gaussian membership functions associated with key input variables and the model target (SI) with respect to the ANFIS model.

reduces the complexity of the ANFIS structure, and provide reliable and accurate estimations of fish diversity in the study area. The proposed machine learning methodology plays a key role in data-mining eco-hydrological applications and provides more information for decision-makers to experience the effects of environmental and water quality variables on the river ecosystems. Future research can keep abreast of the variations in river ecosystems based on this hybrid framework, which can provide precise information to decision-makers for planning the sustainable river ecosystems.

Acknowledgment

This study was funded by the Water Resources Planning Institute, Water Resources Agency, Taiwan (Grant No. MOEAWRA1040202), and the Ministry of Science and Technology, Taiwan (Grant No. 103-2313-B-002-016-MY3). The data provided by the Environmental Protection

Administration (EPA), the Taipei Water Management Office, and the Water Resources Agency, Taiwan, ROC, are very much appreciated. The authors would like to thank the Editors and anonymous Reviewers for their constructive comments that are greatly contributive to the revision of the manuscript.

References

Acreman, M., Arthington, A.H., Colloff, M.J., Couch, C., Crossman, N.D., Dyer, F., et al., 2014. Environmental flows for natural, hybrid, and novel riverine ecosystems in a changing world. Front. Ecol. Environ. 12 (8), 466–473.

Allouche, O., Kalyuzhny, M., Moreno-Rueda, G., Pizarro, M., Kadmon, R., 2012. Area-heterogeneity tradeoff and the diversity of ecological communities. Proc. Natl. Acad. Sci. U.S.A. 109 (43), 17495–17500.

Arthington, A.H., Bunn, S.E., Poff, N.L., Naiman, R.J., 2006. The challenge of providing environmental flow rules to sustain river ecosystems. Ecol. Appl. 16 (4), 1311–1318.

Arthington, A.H., Naiman, R.J., McClain, M.E., Nilsson, C., 2010. Preserving the biodiversity and ecological services of rivers: new challenges and research opportunities. Freshw. Biol. 55 (1), 1–16. https://doi.org/10.1111/j.1365-2427.2009.02340.x.

Arthington, A.H., Dulvy, N.K., Gladstone, W., Winfield, I.J., 2016. Fish conservation in freshwater and marine realms: status, threats and management. Aquat. Conserv. Mar. Freshw. Ecosyst. 26 (5), 838–857.

Barzegar, R., Moghaddam, A.A., Deo, R., Fijani, E., Tziritis, E., 2018. Mapping groundwater contamination risk of multiple aquifers using multi-model ensemble of machine learning algorithms. Sci. Total Environ. 621, 697–712.

Blanes-Vidal, V., Cantuaria, M.L., Nadimi, E.S., 2017. A novel approach for exposure assessment in air pollution epidemiological studies using neuro-fuzzy inference systems: comparison of exposure estimates and exposure-health associations. Environ. Res. 154, 196–203.

Chang, F.J., Tsai, M.J., 2016. A nonlinear spatio-temporal lumping of radar rainfall for modelling multi-step-ahead inflow forecasts by data-driven techniques. J. Hydrol. 535, 256–269.

Chang, F.J., Tsai, Y.H., Chen, P.A., Coynel, A., Vachaud, G., 2015. Modeling water quality in an urban river using hydrological factors–Data driven approaches. J. Environ. Manag. 151, 87–96.

Chang, F.J., Chen, P.A., Chang, L.C., Tsai, Y.H., 2016. Estimating spatio-temporal dynamics of stream total phosphate concentration by soft computing techniques. Sci. Total Environ. 562, 228–236.

Chang, F.J., Tsai, W.P., Chen, H.K., Yam, R.S.W., Herricks, E.E., 2013. A self-organizing radial basis network for estimating riverine fish diversity. J. Hydrol. 476, 280–289.

Chang, F.J., Tsai, M.J., Tsai, W.P., Herricks, E.E., 2008. Assessing the ecological hydrology of natural flow conditions in Taiwan. J. Hydrol. 354 (1–4), 75–89.

Chang, F.J., Tsai, W.P., Wu, T.C., Chen, H.K., Herricks, E.E., 2011. Identifying natural flow regimes using fish communities. J. Hydrol. 409, 328–336.

Chang, L.C., Amin, M., Yang, S.N., Chang, F.J., 2018. Building ANN-based regional multi-step-ahead flood inundation forecast models. Water 10 (9), 1283.

Chang, N.B., Bai, K., Chen, C.F., 2017. Integrating multisensor satellite data merging and image reconstruction in support of machine learning for better water quality management. J. Environ. Manag. 201, 227–240.

Chang, Y.T., Chang, L.C., Chang, F.J., 2005. Intelligent control for modeling of real time reservoir operation, part II: artificial neural network with operating rule curves. Hydrol. Process.: Int. J. 19 (7), 1431–1444.

Chen, I.S., 2009. Indicator Species of Riverine Fishes in Taiwan. Primary Freshwater Fishes, vol 1. National Taiwan Ocean University Press, Keelung, pp. 135.

Chen, I.S., 2011. Survey and Conservation Strategy of the Freshwater Fish Resources in Rivers, Lakes and Ponds of Taiwan (2). Council of Agriculture, Forestry Bureau Press, Taipei, pp. 452.

Chen, I.T., Chang, L.C., Chang, F.J., 2018. Exploring the spatio-temporal interrelation between groundwater and surface water by using the self-organizing maps. J. Hydrol. 556, 131–142.

Cheng, S.T., Herricks, E.E., Tsai, W.P., Chang, F.J., 2016. Assessing the natural and anthropogenic influences on basin-wide fish species richness. Sci. Total Environ. 572, 825–836.

Cheng, S.T., Tsai, W.P., Yu, T.C., Herricks, E.E., Chang, F.J., 2018. Signals of stream fish homogenization revealed by AI-based clusters. Sci. Rep. 8 (1), 15960.

Destouni, G., Fischer, I., Prieto, C., 2017. Water quality and ecosystem management: data-driven reality check of effects in streams and lakes. Water Resour. Res. 53 (8), 6395–6406.

Dou, X., Yang, Y., 2018. Estimating forest carbon fluxes using four different data-driven techniques based on long-term eddy covariance measurements: model comparison and evaluation. Sci. Total Environ. 627, 78–94.

Fashola, M., Ngole-Jeme, V., Babalola, O., 2016. Heavy metal pollution from gold mines: environmental effects and bacterial strategies for resistance. Int. J. Environ. Res. Publ. Health 13 (11), 1047.

Forio, M.A.E., Mouton, A., Lock, K., Boets, P., Nguyen, T.H.T., Ambarita, M.N.D., et al., 2017. Fuzzy modelling to identify key drivers of ecological water quality to support decision and policy making. Environ. Sci. Pol. 68, 58–68.

Förstner, U., Wittmann, G.T., 2012. Metal Pollution in the Aquatic Environment. Springer Science & Business Media.

Gillespie, B.R., Desmet, S., Kay, P., Tillotson, M.R., Brown, L.E., 2015. A critical analysis of regulated river ecosystem responses to managed environmental flows from reservoirs. Freshw. Biol. 60 (2), 410–425.

Goyal, M.K., Bharti, B., Quilty, J., Adamowski, J., Pandey, A., 2014. Modeling of daily pan evaporation in sub tropical climates using ANN, LS-SVR, Fuzzy Logic, and ANFIS. Expert Syst. Appl. 41 (11), 5267–5276.

Halgamuge, M.N., Davis, D., 2019. Lessons learned from the application of machine learning to studies on plant response to radio-frequency. Environ. Res. 178, 108634.

Hofmann, J., Karthe, D., Ibisch, R., Schäffer, M., Avlyush, S., Heldt, S., Kaus, A., 2015. Initial characterization and water quality assessment of stream landscapes in northern Mongolia. Water 7 (7), 3166–3205.

Hong, M., Wang, D., Wang, Y., Zeng, X., Ge, S., Yan, H., Singh, V.P., 2016. Mid-and long-term runoff predictions by an improved phase-space reconstruction model. Environ. Res. 148, 560–573.

Hortal, J., Carrascal, L.M., Triantis, K.A., Thébault, E., Meiri, S., Sfenthourakis, S., 2013. Species richness can decrease with altitude but not with habitat diversity. Proc. Natl. Acad. Sci. Unit. States Am. 110 (24), E2149–E2150.

Jang, J.-S.R., 1993. ANFIS: adaptive-network-based fuzzy inference system. IEEE Trans. Syst.,Man, Cybernetics 23, 665–685.

Jones, F.C., Plewes, R., Murison, L., MacDougall, M.J., Sinclair, S., Davies, C., et al., 2017. Random forests as cumulative effects models: a case study of lakes and rivers in Muskoka, Canada. J. Environ. Manag. 201, 407–424.

Kaab, A., Sharifi, M., Mobli, H., Nabavi-Pelesaraei, A., Chau, K.W., 2019. Combined life cycle assessment and artificial intelligence for prediction of output energy and environmental impacts of sugarcane production. Sci. Total Environ. 664, 1005–1019.

Kail, J., Arle, J., Jahnig, S.C., 2012. Limiting factors and thresholds for macroinvertebrate assemblages in European rivers: empirical evidence from three datasets on water quality, catchment urbanization, and river restoration. Ecol. Indicat. 18, 63–72.

Keylock, C.J., 2005. Simpson diversity and the Shannon–Wiener index as special cases of a generalized entropy. Oikos 109 (1), 203–207.

Kwon, Y.S., Li, F., Chung, N., Bae, M.J., Hwang, S.J., Byoen, M.S., Park, Y.S., 2012. Response of fish communities to various environmental variables across multiple spatial scales. Int. J. Environ. Res. Publ. Health 9 (10), 3629–3653.

Lawton, J.H., MacGarvin, M., Heads, P.A., 1987. Effects of altitude on the abundance and species richness of insect herbivores on bracken. J. Anim. Ecol. 147–160.

Lee, C.S., Lee, Y.C., Chiang, H.M., 2016. Abrupt state change of river water quality (turbidity): effect of extreme rainfalls and typhoons. Sci. Total Environ. 557, 91–101.

Lindman, H.R., 1974. Analysis of Variance in Complex Experimental Designs. San Francisco: W. H. Freeman & Co, Hillsdale, NJ USA: Erlbaum.

Liu, Z., Huang, S., Sun, G., Xu, Z., Xu, M., 2012. Phylogenetic diversity, composition and distribution of bacterioplankton community in the Dongjiang River, China. FEMS Microbiol. Ecol. 80 (1), 30–44.

Marzin, A., Verdonschot, P.F., Pont, D., 2013. The relative influence of catchment, riparian corridor, and reach-scale anthropogenic pressures on fish and macroinvertebrate assemblages in French rivers. Hydrobiologia 704 (1), 375–388.

Meng, W., Zhang, N., Zhang, Y., et al., 2009. Integrated assessment of river health based on water quality, aquatic life and physical habitat. J. Environ. Sci. 21 (8), 1017–1027.

Milliman, J.D., Lee, T.Y., Huang, J.C., Kao, S.J., 2017. Impact of catastrophic events on small mountainous rivers: temporal and spatial variations in suspended-and dissolved-solid fluxes along the Choshui River, central western Taiwan, during typhoon Mindulle, July 2–6, 2004. Geochem. Cosmochim. Acta 205, 272–294.

Moerke, A.H., Lamberti, G.A., 2006. Scale-dependent influences on water quality, habitat, and fish communities in streams of the Kalamazoo river basin, Michigan (USA). Aquat. Sci. 68 (2), 193–205.

Moghaddamnia, Ghafari-Gousheh, M., Piri, J., Amini, S., Han, D., 2008. Evaporation estimation using artificial neural networks and adaptive neuro-fuzzy inference system techniques. Adv. Water Resour. 10, 1016.

Mount, N.J., Maier, H.R., Toth, E., Elshorbagy, A., Solomatine, D., Chang, F.J., Abrahart, R.J., 2016. Data-driven modelling approaches for social-hydrology: opportunities and challenges within the panta rhei science plan. Hydrol. Sci. J. 61 (7), 1192–1208.

Nabavi-Pelesaraei, A., Rafiee, S., Mohtasebi, S.S., Hosseinzadeh-Bandbafha, H., Chau, K.W., 2018. Integration of artificial intelligence methods and life cycle assessment to predict energy output and environmental impacts of paddy production. Sci. Total Environ. 631, 1279–1294.

Nieto, P.G., Fernández, J.A., de Cos Juez, F.J., Lasheras, F.S., Muñiz, C.D., 2013. Hybrid modelling based on support vector regression with genetic algorithms in forecasting the cyanotoxins presence in the Trasona reservoir (Northern Spain). Environ. Res. 122, 1–10.

Nilsson, C., Renöfält, B.M., 2008. Linking flow regime and water quality in rivers: a challenge to adaptive catchment management. Ecol. Soc. 13 (2).

Noori, R., Karbassi, A.R., Moghaddamnia, A., Han, D., Zokaei-Ashtiani, M.H., Farokhnia, A., Gousheh, M.G., 2011. Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. J. Hydrol. 401 (3–4), 177–189.

Noori, R., Deng, Z., Kiaghadi, A., Kachoosangi, F.T., 2015. How reliable are ANN, ANFIS, and SVM techniques for predicting longitudinal dispersion coefficient in natural rivers? J. Hydraul. Eng. 142 (1), 04015039.

Olden, J.D., Naiman, R.J., 2010. Incorporating thermal regimes into environmental flows assessments: modifying dam operations to restore freshwater ecosystem integrity. Freshw. Biol. 55 (1), 86–107.

Papadaki, C., Soulis, K., Muñoz-Mas, R., Martinez-Capel, F., Zogaris, S., Ntoanidis, L., Dimitriou, E., 2016. Potential impacts of climate change on flow regime and fish habitat in mountain rivers of the south-western Balkans. Sci. Total Environ. 540, 418–428.

Piperac, M.S., Milošević, D., Simić, S., Simić, V., 2016. The utility of two marine community indices to assess the environmental degradation of lotic systems using fish communities. Sci. Total Environ. 551, 1–8.

Remesan, R., Shamim, M.A., Han, D., 2008. Model data selection using gamma test for daily solar radiation estimation. Hydrol. Process. 22, 4301–4309.

Remesan, R., Shamim, M.A., Han, D., Mathew, J., 2009. Runoff prediction using an

integrated hybrid modelling scheme. J. Hydrol. 372 (1–4), 48–60.
Sannigrahi, S., Chakraborti, S., Joshi, P.K., Keesstra, S., Sen, S., Paul, S.K., et al., 2019. Ecosystem service value assessment of a natural reserve region for strengthening protection and conservation. J. Environ. Manag. 244, 208–227.
Schinegger, R., Trautwein, C., Melcher, A., Schmutz, S., 2012. Multiple human pressures and their spatial patterns in European running waters. Water Environ. J. 26, 261–273.
Segurado, P., Branco, P., Jauch, E., Neves, R., Ferreira, M.T., 2016. Sensitivity of river fishes to climate change: the role of hydrological stressors on habitat range shifts. Sci. Total Environ. 562, 435–445.
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (379–423), 623–656.
Shi, B., Wang, P., Jiang, J., Liu, R., 2018. Applying high-frequency surrogate measurements and a wavelet-ANN model to provide early warnings of rapid surface water quality anomalies. Sci. Total Environ. 610, 1390–1399.
Simmler, M., Suess, E., Christl, I., Kotsev, T., Kretzschmar, R., 2016. Soil-to-plant transfer of arsenic and phosphorus along a contamination gradient in the mining-impacted Ogosta River floodplain. Sci. Total Environ. 572, 742–754.
Spellerberg, I.F., Fedor, P.J., 2003. A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon–Wiener'Index. Global Ecol. Biogeogr. 12 (3), 177–179.
Stefánsson, A., Koncar, N., Jones, A.J., 1997. A note on the Gamma test. Neural Comput. Appl. 5, 131–133.
Suen, J.P., Eheart, J.W., 2006. Reservoir management to balance ecosystem and human needs: incorporating the paradigm of the ecological flow regime. Water Resour. Res. 42 (3), W03417. https://doi.org/10.1029/2005WR004314.
Suen, J.P., Herricks, E.E., 2006. Investigating the causes of fish community change in the Tahan River (Taiwan) using an autecology matrix. Hydrobiologia 568 (1), 317–330.
Tian, J., Li, C., Liu, J., Yu, F., Cheng, S., Zhao, N., Wan Jaafar, W.Z., 2016. Groundwater depth prediction using data-driven models with the assistance of gamma test. Sustainability 8 (11), 1076.
Tsai, W.P., Chang, F.J., Herricks, E.E., 2016. Exploring the ecological response of fish to flow regime by soft computing techniques. Ecol. Eng. 87, 9–19.
Tsai, W.P., Chang, F.J., Chang, L.C., Herricks, E.E., 2015. AI techniques for optimizing multi-objective reservoir operation upon human and riverine ecosystem demands. J. Hydrol. 530, 634–644.
Tsai, W.P., Huang, S.P., Cheng, S.T., Shao, K.T., Chang, F.J., 2017. A data-mining framework for exploring the multi-relation between fish species and water quality through self-organizing map. Sci. Total Environ. 579, 474–483.
Uen, T.S., Chang, F.J., Zhou, Y., Tsai, W.P., 2018. Exploring synergistic benefits of Water-Food-Energy Nexus through multi-objective reservoir optimization schemes. Sci. Total Environ. 633, 341–351.
Varma, S., Simon, R., 2006. Bias in error estimation when using cross-validation for model selection. BMC Bioinf. 7 (1), 91.
Whitehead, P.G., Wilby, R.L., Battarbee, R.W., Kernan, M., Wade, A.J., 2009. A review of the potential impacts of climate change on surface water quality. Hydrol. Sci. J. 54 (1), 101–123.
Woznicki, S.A., Nejadhashemi, A.P., Abouali, M., Herman, M.R., Esfahanian, E., Hamaamin, Y.A., Zhang, Z., 2016. Ecohydrological modeling for large-scale environmental impact assessment. Sci. Total Environ. 543, 274–286.
Xu, E.G., Bui, C., Lamerdin, C., Schlenk, D., 2016. Spatial and temporal assessment of environmental contaminants in water, sediments and fish of the Salton Sea and its two primary tributaries, California, USA, from 2002 to 2012. Sci. Total Environ. 559, 130–140.
Yaseen, Z.M., Ebtehaj, I., Bonakdari, H., Deo, R.C., Mehr, A.D., Mohtar, W.H.M.W., et al., 2017. Novel approach for streamflow forecasting using a hybrid ANFIS-FFA model. J. Hydrol. 554, 263–276.
Zhao, C., Zhang, Y., Yang, S., Xiang, H., Sun, Y., Yang, Z., et al., 2018. Quantifying effects of hydrological and water quality disturbances on fish with food-web modeling. J. Hydrol. 560, 1–10.
Zhou, Y., Guo, S., Chang, F.J., 2019. Explore an evolutionary recurrent ANFIS for modelling multi-step-ahead flood forecasts. J. Hydrol. 570, 343–355.