ISEIS

Journal of Environmental Informatics Letters 1(1) 37-47 (2019)

Journal of Environmental Informatics Letters

www.iseis.org/jeil

Ensemble Learning Enhanced Stepwise Cluster Analysis for River Ice Breakup Date Forecasting

W. Sun[1 *], Q. Shi[1 **], Y. Huang[2], and Y. Lv[3]

[1] *School of Geography and Planning, Sun Yat-Sen University, Guangzhou, Guangdong 510275, China*
[2] *School of Civil and Environmental Engineering, Nanyang Technological University, 50 Nanyang Avenue 639798, Singapore*
[3] *Ministry of Education Key Laboratory for Transportation Complex Systems Theory and Technology, School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, China*

Received 21 January 2019; revised 13 February 2019; accepted 10 March 2019; published online 31 March 2019

ABSTRACT. Frequently occurring ice jams often cause concern in northern regions. Breakup timing is directly related to emergency responses preparation and thus its early accurate forecasting is beneficial to ice-related flooding management. The stepwise cluster analysis (SCA) is a non-parameter regression method, which generates a classification tree in the sense of probability through cutting or merging operations according to certain statistic criteria. To enhance SCA's predictive performance, a SCA ensemble (SCAE) method is developed and applied to forecasting of annual river ice breakup dates (BDs). In detail, the SCA is employed as a base model at the lower level while the simple average method is selected as combining models at the upper level. The SCA base models are selected according to different performance selection criteria and searched for further combination. A site on a representative river prone to river ice flooding in Alberta, Canada is selected to demonstrate the effectiveness of the proposed SCAE. The results mainly show that: the SCA base models with multiple combinations of inputs and internal parameters are able to predict the BDs with good performances (the highest average of correlation coefficients for training can be 0.958); the optimal SCA base model has three inputs, which indicates that the temperatures before breakup and just after freeze-up as well as the maximum of water flow in March are relatively important indicators of BD. The optimal SCAE, including base models from different performance selection criteria, has the lowest average of root mean squared error, which improves upon the optimal SCA base model by 25.3%. It indicates the different model selection criteria do improve the diversity and thus further help to improve the performance of ensemble models. This first application of the SCAE to river ice forecasting highlights the possibility of using the ensemble learning paradigm to enhance the SCA. The potential applications of the SCAE to other forecasting problems are expected.

Keywords: river ice, breakup, stepwise cluster analysis, ensemble learning

1. Introduction

Long-term river flow patterns are affected by climate change and anthropogenic stressors (e.g. flow regulation and water usage). As an important annual hydrologic event in each spring, northern river ice breakup occurs due to warming temperatures and rising water flows (Beltaos, 2000; Beltaos and Prowse, 2009). When break-up conditions favor the development of ice jams, sudden extreme flooding may occur at certain reaches. The river ice break-up forecasting models are useful tools to support flooding management. Among the forecasted variable of interest, breakup timing is directly related to emergency responses preparation and thus its early accurate forecasting is beneficial to ice-related flooding management. Previously, relatively limited data-driven models, such as empirical equations (Gao et al., 2012), artificial neural networks (Wang et al., 2008; Zhao et al., 2012) and support vector machines (Zhou et al., 2009) were developed for river ice timing forecasting. This is because the river ice forecasting task is a challenge due to the complicated interactions between weather and upstream-downstream river ice conditions (Mahabir et al., 2006; Hicks, 2009). Furthermore, it would be more challenging because of scarce ice data and inherent uncertainty caused by the limited and difficult observation conditions.

The stepwise cluster analysis (SCA) is a non-parameter regression method. In SCA, cutting or merging operations according to certain statistic criteria are implemented to generate a classification tree in the sense of probability. The SCA tree is able to explicitly describe nonlinear relationships among continuous and/or discrete variables. Nevertheless, besides its calculation complexity, the performance of the SCA is sensitive to its inputs and internal parameters; the difference within leaf clusters of a SCA tree is usually not well described. The SCA has been applied to various water resources and environmental management problems, such as urban air quality prediction (Huang, 1992), lung cancer diagnosis (Ren et al., 1997), waste

* Corresponding author. Tel.: +86 02084112834.
E-mail address: sunwei29@mail.sysu.edu.cn (W. Sun).
** Corresponding author.
E-mail address: shixi5@mail.sysu.edu.cn (Q. Shi).

ISSN: 2663-6859 print/2663-6867 online
 doi:10.3808/jeil.201900005

treatment process simulation (Sun et al., 2009; Sun et al., 2011), groundwater bioremediation optimization (Huang et al., 2006; Qin et al., 2007; He et al., 2008b; Wang et al., 2012; Zhao et al., 2017), open water forecasting (Fan et al., 2015; Li et al., 2015; Han et al., 2016; Zhuang et al., 2016b; Cheng et al., 2016; Fan et al., 2017;) and climate model downscaling (Wang et al., 2013; Zhuang et al., 2016a; Zhai et al., 2019). However, few applications of SCA to river ice forecasting are reported. Meanwhile, it has been reported that stacking ensemble learning can improve the overall performance through effective combinations of different base models (Erdal and Karakurt, 2013; Sun and Trevor, 2017; Sun and Trevor, 2018a). Thus, further improving the accuracy of SCA predictability using the stacking ensemble learning paradigm and its application to river ice forecasting are desired.

The purpose of this study is to develop a method of ensemble learning enhanced stepwise cluster analysis or SCA ensemble (SCAE) and to apply it to river ice breakup date forecasting. It will entail: (i) the SCA with multiple inputs will be developed and selected as base models according to different performance criteria; (ii) the outputs of several SCA base models will be combined as inputs of the SCA ensemble models; and (iii) the proposed SCAE will be applied to a representative site in an unregulated river in Alberta, Canada, where frequent ice-caused flooding is a concern. This study would benefit operational river-ice related flooding management though demonstration of improving the SCA models based on the stacking ensemble learning paradigm.

2. Methods

2.1. Stepwise Cluster Analysis (SCA)

The SCA has the potential to be developed as a forecasting model for river ice break-up dates. From the perspective of SCA, the whole or part of the training set is a single cluster (α), including n_α samples, m independent variables (X), and one dependent variable (Y). The cluster α can be cut into two sub-clusters β and γ (with n_β and n_γ samples, respectively). Based on the Wilks' likelihood-ratio criterion, the cutting point is optimal only if the value of Wilks' Λ is minimum (Wilks, 1962). In the sense of Wilk's likelihood-ratio criterion, the smaller the Λ value, the larger the difference between the sample means of β and γ. Because the Λ is directly associated with the F statistic (Rao, 1952), the sample averages of the two sub-clusters can be evaluated for significant differences through the F test. Thus, the criteria of cutting (or merging or not) clusters are transformed to making a set of F tests (Rao, 1965; Tatsuoka, 1971). All sub-clusters derived from the original cluster (α) will enter a set of iterative runs of cutting (or merging) processes until either all hypotheses of further cut (or mergence) operations are rejected or the minimum number of samples (N_{min}) within every cluster is satisfied. Once all calculations and tests are completed, a SCA tree can be generated which indicates the training is over.

The SCA is more similar to a model tree, such as a CART regression tree model (Breiman et al., 1984), instead of clustering algorithms, since the dependent variable (predicted value of interest) should be provided. The final structure of SCA is similar to that a regression tree but with more statistical meanings. The value of the dependent variable at each leaf node is based on a series of splitting rules of independent variables at the branching nodes (De'ath and Fabricius, 2000; Erdal and Karakurt, 2013; Galelli and Castelletti, 2013). These rules in SCA are based on the Wilk's likelihood-ratio criterion, which is different from the total residual sum square estimation used in the CART. Once the structure of SCA is calibrated by the training set, the values of the independent variables of new samples will be used to determine which leaf node a sample enters. When entering the tree, a new sample ($x_1, x_2, \ldots, x_m, y_1$; y_1 is unknown) will finally drop into a leaf cluster which can be neither cut nor merged further. The routes from top to bottom are decided by a comparison between new independent variables ($x_1, x_2, \ldots, x_m$) and the corresponding threshold at each branching node. The predicted value of y_1 will be the average of dependent variables of the training samples in the leaf cluster. Thus, the SCA tree is able to predict new dependent variables once new samples enter the tree. A more detailed description of SCA can be referred to the previous work (Qin et al., 2007). An open-source R package (rSCA) is developed for stepwise cluster analysis, which is available without charge at http://cran. rproject.org/package = rSCA (Wang et al., 2015).

2.2. SCA Ensemble

Ensemble learning is a machine learning paradigm which combines multiple learners to solve the same problem (Dietterich, 2000). These learners in an ensemble are usually called base learners (models). The generalization ability of an ensemble is usually better than that of a single learner, as long as two necessary conditions of base learners are satisfied: accuracy and diversity. In other words, base learners with more diversity and better accuracy will be beneficial to the performance improvement of the ensemble learning (Polikar, 2006; Wang et al., 2011). As one of popular ensemble learning methods, stacking involves a higher-level (metalevel) learner (combining model) to combine lower-level (base-level) learners (base model) to achieve greater predictive performance (Wolpert, 1992; Ting and Witten, 1999). In this sense, stacking is similar to the multiple model combination method (Sun and Trevor, 2018a). Although stacking is usually employed to combine different-type base learners built by multiple learning algorithms, it can be used to combine same-type base learners with different structures or calibrated parameters as well.

In this study, the SCAE for annual river ice breakup dates has a two-level structure, which includes base and combining models (Figure 1). In terms of its functions, the base models link the BDs with their corresponding indicators; the combining models build the relations between the predicted BDs by each base model and the observed BDs. Since the inputs represent different information to be used in the models, the SCA models with different inputs can be developed as base models with performance diversity and proper accuracy. The simple average method (SAM) is selected as the combining model since it is widely used with demonstrated performance improvement and

simplicity for implementation.

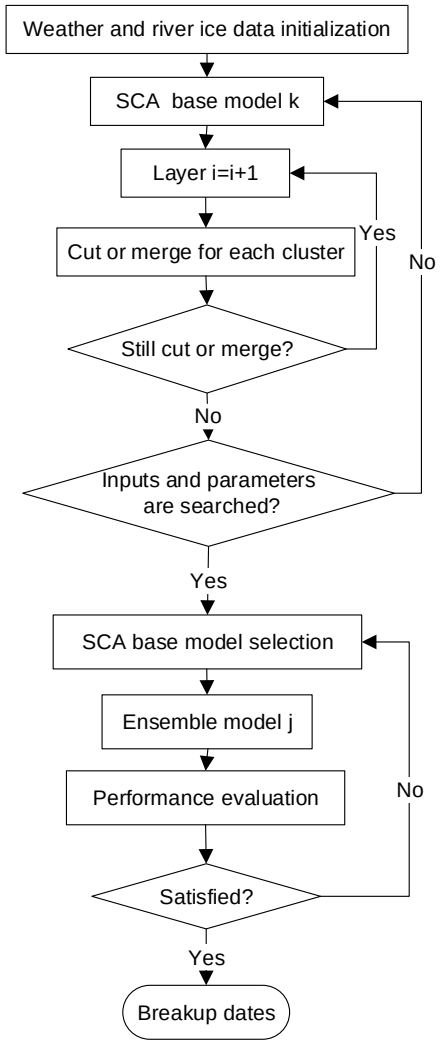


Figure 1. Flowchart of the SCAE model.

2.3. Model Development

To determine the structure of a stacking ensemble model, the first step is to develop the set of base-level learners though applying proper learning algorithms based on the original dataset or its subsets; The second step of stacking is to collect both the output of each base learner and the output in the original dataset, and then combine them into a new dataset. In other words, the output of each base learner will be the inputs in the new dataset while the output is the same. The final step is to construct the meta-level learners using the new dataset.

Since most of the prediction models can often encounter the over-fitting problem, separate verification of the validation set can help maintain the generality of the calibrated models. In addition, considering the river ice data set is scarce, the leave-one-out cross validation (LOOCV) is accordingly employed to evaluate the generality of the proposed SCA base models. For the data set with *m* samples, each SCA base model is calibrated with *m* - 1 samples and validated with 1 sample reserved for each run of LOOCV. The implementation of all *m* runs ensures all samples be selected in the validation sets once. Similarly, the output of select SCA base models is employed as the inputs of the SAM models The SAM model will be calibrated and validated by the LOOCV as well.

To evaluate the predictive performance of the proposed member and combining models for river ice break-up dates, two performance indices (correlation coefficient (*R*) and root mean squared error (RMSE)) were selected. The model's performance becomes better with higher *R* (closer to one) and lower RMSE (closer to zero). The evaluation indices are expressed as follows:

$$R = \frac{\sum_{j=1}^{n}(Y_{sj} - \bar{Y}_s)(Y_j - \bar{Y})}{\sqrt{\sum_{j=1}^{n}(Y_{sj} - \bar{Y}_s)^2}\sqrt{\sum_{j=1}^{n}(Y_j - \bar{Y})^2}} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum_{j=1}^{n}(Y_j - Y_{sj})^2}{n-1}} \quad (2)$$

where n is the sample number in the training or validation set, Y_j and Y_{sj} are the observed and predicted BDs in the j_{th} sample, respectively; and $\bar{Y}$ and $\bar{Y}_s$ are the mean of the observed and predicted BDs. In LOOCV, these evaluation indices base on multiple training and validation sets will be averaged to reflect the overall performance of the proposed SCAE methods.

3. Application

The study area is located at the confluence between the Athabasca and Clearwater Rivers in the town of Fort Mc-Murray in Alberta, Canada (Figure 2). The Athabasca River, with its basin covering 95,300 km^2, is the largest unregulated river in Alberta. It originates in the Columbia Ice Fields and flows 1,231 km across central and northern Alberta before discharging into the Peace-Athabasca delta. Fort Mc-Murray is located approximately 900 river kilometers downstream of the Athabasca River's glacial headwaters. The about 200 km reach upstream of Fort Mc-Murray is characterized by numerous rapids and bed discontinuities (Kowalczyk and Hicks, 2003; Sun et al., 2015). When river ice breakup occurs, a cascade of small ice jams or ice accumulations in this reach may release and progress downstream. In the 2.5 km reach downstream of the Grant MacEwan Bridge in Fort McMurray, the river widens and forms numerous islands and bars, while the river slope decreases by an order of magnitude (She et al., 2009; Andrishak and Hicks, 2011). Additionally, the Clear-water River discharges into the Athabasca River at this location. These factors may result in possible ice jam formation during river ice breakup, which increases the flooding risk in the town.

The historical BDs at this location from 1980 to 2015 was collected from the official website of Regional Municipality of Wood Buffalo (http://www.rmwb.ca/). These break-up dates are actually the last date during the breakup process, which

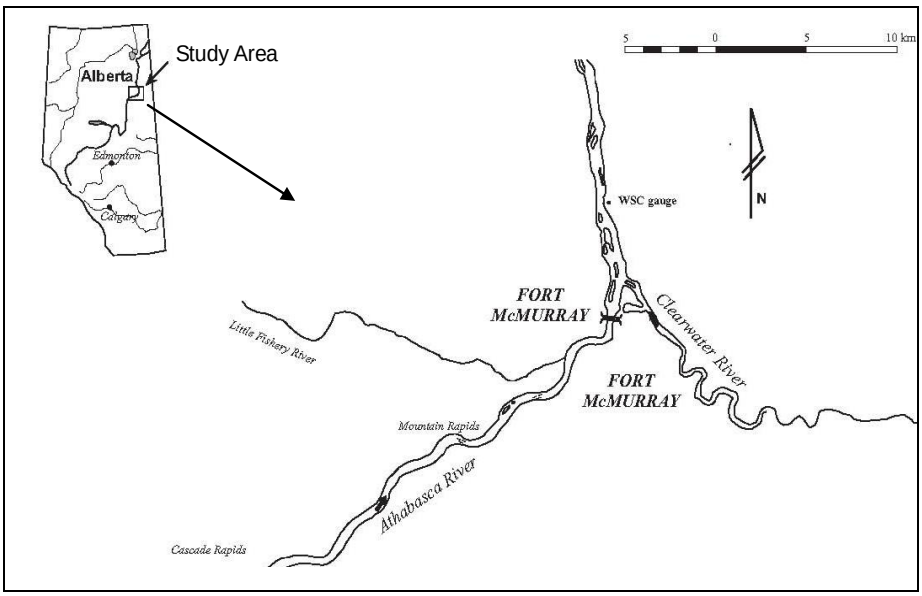


Figure 2. The Athabasca River at Fort McMurray in Alberta, Canada.

implies open channel conditions or minimal ice-related flooding risks at the town. This definition is more meaningful in the viewpoint of emergency management in response to the lasting flooding potential near to and during breakup. Although this definition is used in this study, the first date of breakup processes (i.e. the first significant movement of ice cover) would be another definition elsewhere (Zhao et al., 2010). Figure 3 shows the distribution of the BDs, which ranges from early April (Julian day: 100) to early May (123). The approximate linear pattern indicates that the BDs follow a normal distribution, which is assumed and required in the SCA model. The potential weather and river indicators corresponding to BDs are employed as the model inputs. The weather indicators are employed based on monthly statistics for daily maximum, minimum and average temperatures of two Environment Canada stations at Fort Mc-Murray (WMO ID: 71689 and 71585); the missing data at station 71689 were complemented by the data from station 71585. The river indicators are based on water flows at the Water Survey Canada gauge of Athabasca River below McMurray (Station 07DA001).

4. Results Analysis

4.1. Selection of Indicators Affecting BDs

Many factors affecting river ice break-up dates include spring surface air temperatures and downstream/upstream river ice conditions (Bieniek et al., 2011; Beltaos and Burrell, 2015; Cooley and Pavelsky, 2016). These indicators are prescreened as the candidate inputs (independent variables) of the models based on two criteria: both availability before river ice breakup and linear correlation coefficients with BDs. It is noted that although the correlation should be nonlinear in nature, the linear correlation coefficients are still helpful to narrow down the range of these indicators. The select candidate climate and river flow indicators corresponding to BDs are listed in Table 1. Among the input variables, X_1 to X_9 are indicators associated with monthly characteristics of daily temperatures at Fort Mc-Murray in last fall or this winter. X_{10} to X_{17} are indicators related to daily water flow at Athabasca River below Mc-Murray during different periods. Notably, the number of those temperature and water flow indicators in March has the maximum values which indicate the weather and river conditions just before breakup may have significant effects on the breakup timing.

4.2. SCA Base Models with the Best Validation Performance (Criterion 1)

The performance of SCA base models depends on adjustment of several factors, which include data quality, combination of input variables, internal configuration parameters (e.g. α_{cut}, α_{merge} and N_{min}), and the data partition strategy. The criteria for cluster analysis are: cutting cluster when $p \leq \alpha_{\text{cut}}$ and merging clusters when $p \geq \alpha_{\text{merge}}$, where the p values are significance levels of F-test. Generally, higher α_{cut} (a decreased strictness in the cutting) would result in lower F_1 level (more cutting operations). Similarly, higher α_{merge} (reduced strictness of mergence) would result in higher F_2 level (more merging operations). The N_{min} also affects the scales of the cluster trees since it is used as one of the ending criteria for training the tree. To identify the optimal structure, multiple combinations of inputs and parameters were testified through the greedy search-based LOOCV method. Based on this method, the maximum number of inputs was searched from 2 to 6; the α_{cut} was searched from 0.01 to 0.05; the α_{merge} was searched from α_{cut} to 0.05; and the N_{min} was searched from 2, 5 and 10. Table 2 lists the representative SCA base models with the best validation performance. All of SCA models have good and diverse performances with different combinations of inputs and internal parameters. In terms of validation performance, the SCA_7 with 5 inputs has the lowest

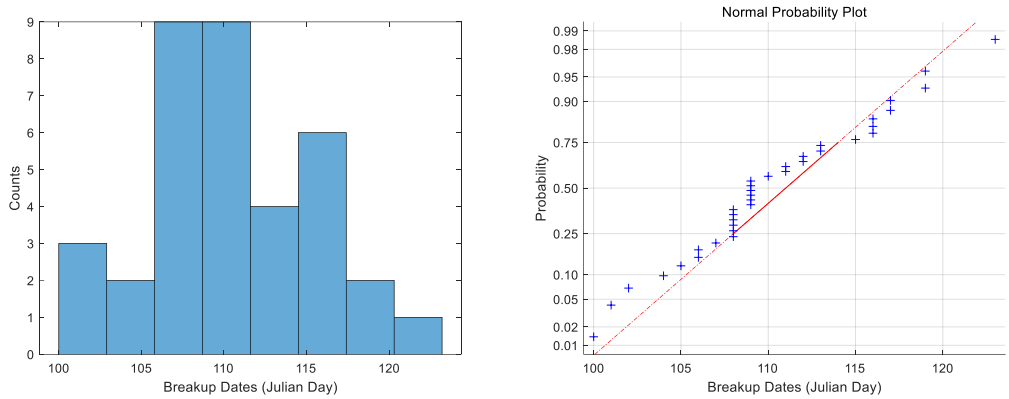


Figure 3. Flowchart of the SCAE model.

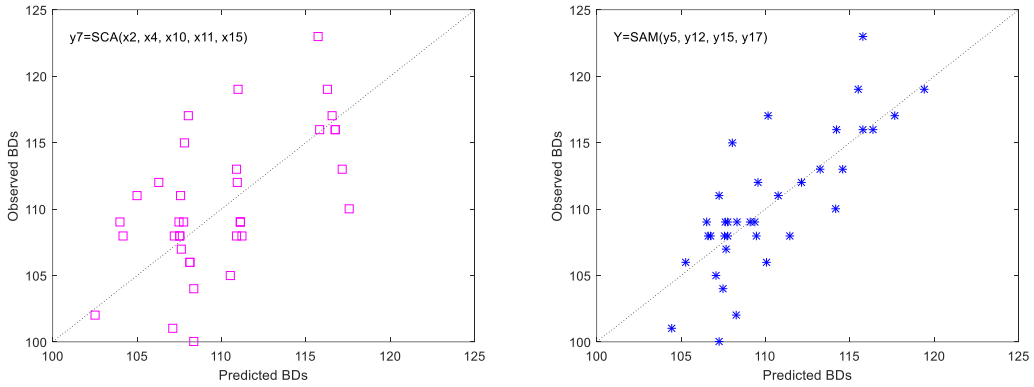


Figure 4. Validation performance comparison of optimal SCA base and ensemble models.

$RMSE_{avg}$, which is considered the optimal SCA base model with the best validation performance. Compared with other base models, the number of inputs in SCA_7 is in the middle range. It implies that an optimal combination of inputs may result in better validation performance and structures of the SCA trees.

4.3. SCA Base Models with Better Training and Validation Performance (Criterion 2)

Table 3 lists representative SCA base models with better training and validation performance when the inputs are set from 2 to 7. The base models with diverse performance are selected for the ensemble framework. For the training performance, the indices (R_{avg} and $RMSE_{avg}$) for the models in Table 3 range from 0.8284 to 0.9579 and from 1.435 to 2.882, respectively. These indices in Table 3 are larger and lower than those in Table 2 ([0.6317, 0.8983], [2.253, 4.003]), respecttively. In terms of validation performance, the ranges of $RMSE_{avg}$ for the models in Table 3 ([4.434, 4.877]) are slightly worse than those in Table 2 ([4.285, 4.534]). As for structure, the SCA base models tend to have higher α_{cut} and α_{merge} and lower N_{min}, which implies more cutting and merging operations during the training process. Additionally, the models with more number of inputs tend to have better training performance. It implies that more complex structures do increase the fitting ability of these models. In contrary, the models in Table 3 with a middle range of inputs (3 and 4) tend to have better validation performance. It indicates that a balanced structure may be more beneficial to the generality of these models.

4.4. SCA Ensemble Models

Table 4 shows representative ensemble models with higher performance. These SCA base models are selected with two different criteria (i.e. best validation performance vs. better training and validation performance). The outputs of these base models (y_1 to y_{18}) are used as candidate inputs for the ensemble models (SAM). The performance was evaluated through searching all possible combinations of these base models when the input number of the SAM is set from 2 to 5. Through comparing the performance indices, most of these ensemble models generally improve upon the base models in terms of the training performance. The R_{avg} for most of ensemble models are greater than 0.96 while their $RMSE_{avg}$ are less than 1.6. It is not surprised since the ensemble models include more base models so that the overall structure complexity increases. As for the validation performance, the $RMSE_{avg}$ for all of ensemble models are less than 3.50 which is an obvious improvement since the $RMSE_{avg}$ for all SCA base models are greater than 4.28. It is

also noted that $enSAM_3$ (3 base models) to $enSAM_8$ (5 base models) have very similar performance, which indicates the improvement robustness of this stacking ensemble learning paradigm. Among all ensemble models, the $enSAM_5$ (4 inputs: y_5, y_{12}, y_{15} and y_{17}) has the lowest $RMSE_{avg}$ (3.200) for validation and very similar R_{avg} (0.9643) and RMSE (1.464) for training to other ensemble models. Since it also has a middle range of structure (4 inputs), the $enSAM_5$ is considered the optimal stacking ensemble model.

4.5. Comparison between the Optimal SCA Base and Ensemble Models

Figure 4 shows the validation performance comparison between the optimal SCA base and ensemble models to further illustrate their ability to predict new samples. The scattered points for the $enSAM_5$ (ensemble model) are generally more aggregated than those for the SCA_7 (base model); the difference between the predicted BDs of the SCA_7 and the observed BDs is relatively larger in different ranges than those of the $enSAM_5$; the $RMSE_{avg}$ for the $enSAM_5$ (3.200) is much lower than that of the SCA_7 (4.285). The differences in scatter plots and performance indices clearly indicate that the former has better validation performance than the latter. In detail, for mid-range predictions, the advantages of the $enSAM_5$ over the SCA_7 are even more noticeable. For the prediction of higher and lower ranges, both $enSAM_5$ and SCA_7 have presented a certain degree of underestimation and overestimation, respectively. The possible reason for the rela tively worse performance of the

Table 1. Candidate Climate and River Flow Indicators

X	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
Variable	T	T	T	T	T	T	T	T	T
Monthly	Max	Max	Min	Avg	Max	Min	Max	Min	Min
Daily	Avg	Avg	Avg	Min	Min	Min	Max	Max	Max
Period	Mar	Last Dec	Last Nov	Mar	Mar	Mar	Mar	Mar	Last Dec
Unit	℃	℃	℃	℃	℃	℃	℃	℃	℃
X	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	
Variable	WL	WL	WL	WL	WL	WL	WL	WL	
Monthly	Avg	Avg	Avg	Max	Max	Min	Min	Min	
Daily	Avg	Avg	Avg	Avg	Avg	Avg	Avg	Avg	
Period	Jan	Mar	Last Nov	Mar	Last Sept	Mar	Last Sept	Last Nov	
Unit	m^3/s	m^3/s	m^3/s	m^3/s	m^3/s	m^3/s	m^3/s	m^3/s	

* T means temperature and WL means water flow

Table 2. Representative SCA Base Models with the Best Validation Performance

Model	Inputs	α-cut	α-merge	N_{min}	Output	Training R_{avg}	$RMSE_{avg}$	Validation $RMSE_{avg}$
SCA_1	x_{10}, x_{15}	0.03	0.05	10	y_1	0.744	3.444	4.466
SCA_2	x_9, x_{10}	0.01	0.01	10	y_2	0.632	4.003	4.534
SCA_3	x_1, x_{10}, x_{15}	0.03	0.04	2	y_3	0.746	3.434	4.409
SCA_4	x_4, x_{10}, x_{15}	0.04	0.04	10	y_4	0.827	2.901	4.454
SCA_5	x_2, x_{10}, x_{11}, x_{15}	0.04	0.05	2	y_5	0.883	2.399	4.318
SCA_6	x_6, x_{10}, x_{11}, x_{15}	0.05	0.05	5	y_6	0.898	2.253	4.364
SCA_7	x_2, x_4, x_{10}, x_{11}, x_{15}	0.03	0.03	10	y_7	0.819	2.971	4.285
SCA_8	x_1, x_2, x_4, x_{10}, x_{15}	0.03	0.03	10	y_8	0.816	2.998	4.397
SCA_9	x_1, x_2, x_4, x_{10}, x_{11}, x_{15}	0.05	0.05	10	y_9	0.845	2.751	4.321
SCA_{10}	x_1, x_6, x_8, x_9, x_{13}, x_{15}	0.01	0.01	2	y_{10}	0.793	3.118	4.469

Table 3. Representative SCA Base Models with Better Training and Validation Performance

Model	Inputs	α-cut	α-merge	N_{min}	Output	Training R_{avg}	$RMSE_{avg}$	Validation $RMSE_{avg}$
SCA_{11}	x_3, x_{15}	0.04	0.04	5	y_{11}	0.828	2.882	4.795
SCA_{12}	x_6, x_{10}, x_{15}	0.04	0.04	10	y_{12}	0.890	2.348	4.486
SCA_{13}	x_2, x_{10}, x_{11}, x_{15}	0.05	0.05	2	y_{13}	0.912	2.081	4.527
SCA_{14}	x_6, x_{10}, x_{11}, x_{15}	0.03	0.03	5	y_{14}	0.907	2.150	4.434
SCA_{15}	x_1, x_2, x_5, x_{11}, x_{16}	0.05	0.05	2	y_{15}	0.905	2.163	4.476
SCA_{16}	x_1, x_2, x_6, x_{10}, x_{11}, x_{15}	0.05	0.05	2	y_{16}	0.912	2.068	4.687
SCA_{17}	x_1, x_2, x_6, x_8, x_{13}, x_{15}	0.05	0.05	2	y_{17}	0.951	1.551	4.795
SCA_{18}	x_1, x_8, x_{11}, x_{12}, x_{13}, x_{15}	0.05	0.05	2	y_{18}	0.958	1.435	4.877

Table 4. Representative SCA Ensemble Models

Model	Inputs	Output	Training R_{avg}	$RMSE_{avg}$	Validation $RMSE_{avg}$
$enSAM_1$	y_{12}, y_{15}	Y_1	0.939	1.826	3.405
$enSAM_2$	y_{15}, y_{17}	Y_2	0.964	1.402	3.495
$enSAM_3$	y_5, y_{15}, y_{17}	Y_3	0.965	1.431	3.237
$enSAM_4$	y_{12}, y_{15}, y_{17}	Y_4	0.967	1.414	3.238
$enSAM_5$*	y_5, y_{12},y_{15}, y_{17}	Y_5	0.964	1.464	3.200
$enSAM_6$	y_5, y_{15}, y_{16}, y_{17}	Y_6	0.966	1.405	3.211
$enSAM_7$	y_5, y_9, y_{12},y_{15}, y_{17}	Y_7	0.961	1.561	3.279
$enSAM_8$	y_5, y_{12}, y_{14}, y_{15}, y_{17}	Y_8	0.960	1.583	3.283

* represents the optimal ensemble model of all models

SCA base model is because the difference within leaf nodes is not handled well, besides the complicated river ice breakup mechanism.

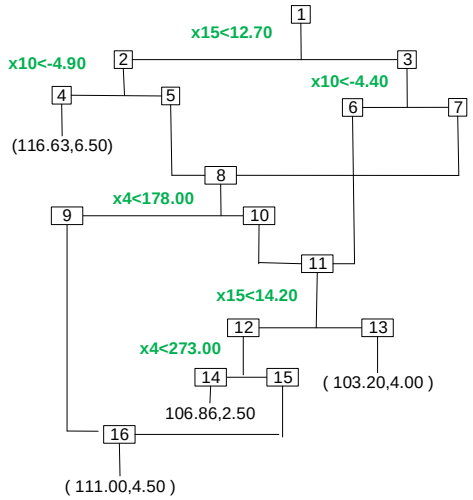


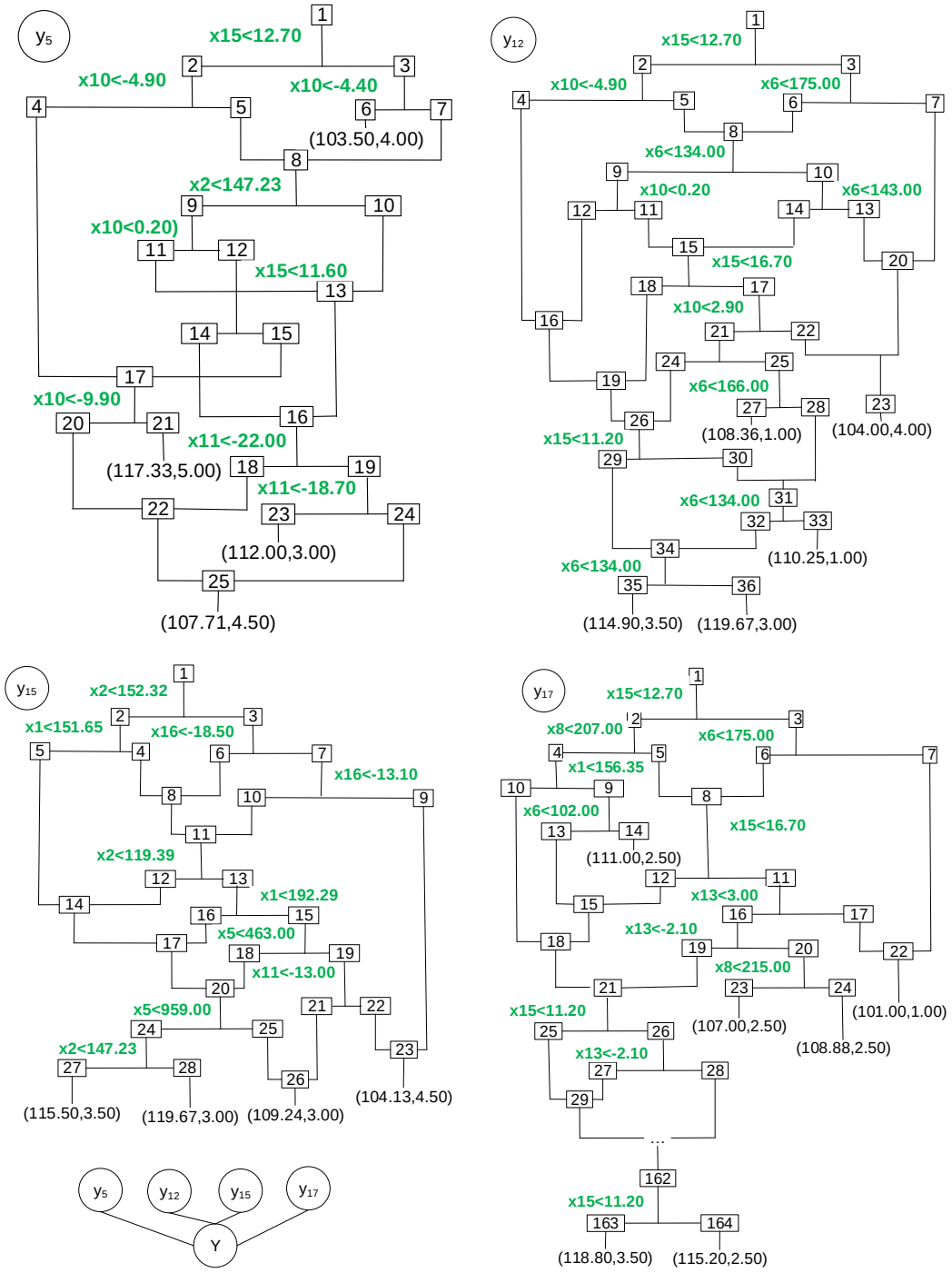
Figure 5. Structure of optimal SCA base model calibrated with all data.

Figure 5 presents the tree structure of the optimal SCA base model (SCA_7). Although the SCA_7 has five inputs (x_2, x_4, x_{10}, x_{11}, and x_{15}) based on the LOOCV, its final structure calibrated by all years of data only has 3 inputs (x_4, x_{10}, and x_{15}). Based on the resulting tree, the BD can be predicted given the inputs of tem- perature and water flow conditions. For example, let x_4 = 273 (m^3/s), x_{10} = -0.7 (℃), and x_{15} = 14.0 (℃) as new inputs. To predict the BD, we have: x_{15} > 12.70 (℃) for the first branch knot so that the sample enters cluster 3 (Figure 5); x_{10} > -4.40 (℃), so that it enters cluster 7 and merges into cluster 8; x_4 > 178.00 (m^3/s), so that it enters cluster 10 and then merges into cluster 11; x_{15} < 14.20 (℃), so that it enters cluster 12; x_4 ≥ 273 (m^3/s) so that it enters cluster 15 and finally merges into cluster 16 with a prediction value of 111.00 ±4.5 (Julian Day). Based on previous SCA studies, the inputs at topper layers have more effects on the output (Sun et al., 2009). Thus, the top branch node is the maximum daily temperatures in March (x_{15}), which implies that the temperatures before breakup do affect the BD to a large extent. The x_{10} (the maximum of daily average temperatures in last December) at the second top layer indicates the temperatures just after freeze-up have certain effects on the BD as well. Finally, the maximum of water flow in March (x_4) is also an important indicator of BD.

Figure 6 illustrates the tree structures of four SCA base models (SCA_5, SCA_{12}, SCA_{15}, and SCA_{17}) within the optimal SCA ensemble model ($enSAM_5$). Noticeably, these four SCA base models have more layers than the optimal SCA base model (SCA_7). Compared with the 3 inputs in the SCA_7, the SCA_5, SCA_{12}, SCA_{15}, SCA_{17} have four (x_2, x_{10}, x_{11}, and x_{15}), three (x_6, x_{10}, and x_{15}), five (x_1, x_2, x_5, x_{11}, and x_{16}), and five (x_1, x_6, x_8, x_{13}, and x_{15}; x_2 is not included in the final structure) inputs, respectively. Among these four models, only SCA_5 has the same selection criteria as SCA_7 since both of them belong to representative base models with the best validation performance (criterion 1); the other models (SCA_{12}, SCA_{15}, and SCA_{17}) are selected due to better training and validation performance (criterion 2). Although the validation performance of SCA_7 is the optimal, the training performance of four base models is better. The R_{avg} of four base models ranges from 0.8831 to 0.9509 while the $RMSE_{avg}$ ranges from 1.551 to 2.399. This indicates that both selection criteria do provide more suitable base models with diverse performance to the framework of the ensemble model and further improve the overall performance.

5. Discussion

Several types of data-driven models have been applied to BD forecasting. The methods developed in the early stages, such as indicators, correlation, and empirical equations, are difficult to transfer to other sites with different river ice regime (Guo, 2002; Gao et al., 2012). The multiple linear regression method is often not appropriate since the linear assumption can hardly be met in river ice forecasting problems. Artificial neural networks (ANNs) and support vector machines (SVMs) are well-accepted tools to quantify the complicated nonlinear river ice breakup relations (Wang et al., 2008; Zhou et al., 2009; Zhao et al., 2012), although these black-box models are relatively difficult to explicitly explain and no statistical meanings can be provided. Compared with these methods, the SCA can establish a nonlinear relation between weather/river indicators and breakup dates, provide significance levels in the cutting or merging steps to control prediction accuracy, and generate the



Fgure 6. Structure of optimal SCA ensemble models.

clustering tree structure describing the inherent logic explicitly.

However, the SCA does have some limitations. Firstly, the original version of SCA can only be applied to the problem where the dependent variables (output) are normally distributed (Huang, 1992). If it is not the case, the output should be transformed to a variable with a normal distribution so that the cutting and merging process can be effectively handled based on the Wilk's statistic. This transformation should be two-directional so that the transformed variable can be processed back to the original output of the SCA model in the final stage. If the output or its transformed form cannot fit the normal distribution, the original version of SCA is not recommended for describing the relation. In this case, other statistics for cutting and merging operations should be introduced to handle the specific

distribution of the output. Secondly, the training for a SCA model involves a set of iterative runs of cutting (or merging) steps to search for all possible clustering possibilities as well as sorting operations according to multiple independent variables. This calculation process is relatively intensive even for a small data set. The implementtation algorithm of the SCA can be optimized based on careful designs to improve the calculation efficiency. Thirdly, similar to other regression tree models, the SCA cannot distinguish the difference within the leaf node. Fortunately, some algorithms have been developed to build a cluster-wise model for each leaf node (He et al., 2008a), which is similar to the M5 model (Quinlan, 1992; Wang and Witten, 1997). Finally, the SCA has a certain ability to filter irrelevant inputs during the training. Overloaded inputs may bring difficulty to the SCA training. Some pre-/post-processing methods or searching algorithms (e.g. genetic algorithm) can also be combined with the SCA to improve its performance (Sun et al., 2011).

Considering the advantages and limitations of the SCA, the proposed ensemble learning framework (the SCAE) provides an innovate manner to enhance the SCA's ability for forecasting river ice BD through combining multiple SCA models. The SCAE is promising based on the demonstrated predictive performance improvement over the optimal SCA base model. However, the performance improvement is associated with an increase of the overall structure complexity (Sun and Trevor, 2018b). Introduction of multiple base models may somehow bring difficulties in revealing breakup mechanisms since diverse tree structure based explanations may exist (Sun, 2018). In terms of the ensemble type, the proposed framework of the SCAE can be categorized as a stacking ensemble. In the future, many other types of ensemble manners, including boosting and bagging (Anctil and Lauzon, 2004; Zaier et al., 2010), can be further applied to the SCA so that the forecasting performance is expected to be further improved. It is also worthy pointing out that the ensemble learning enhanced SCA has the general applicability to various prediction or forecasting problems.

6. Conclusions

A method of SCA ensemble (SCAE) is developed and applied to forecasting of annual river ice breakup dates (BDs). The SCA is employed as base models at the lower level while the simple average method (SAM) is selected as combining models at the upper level. Within the two-level structure of the SCAE, the fitting and generality ability of the SCA can be enhanced by the ensemble learning paradigm. The Athabasca River at Fort Mc-Murray in Alberta, Canada is selected as the study area, where frequently occurring ice jams are a concern. The historical breakup data in 1980 ~ 2015 was employed to evaluate the performance of the proposed SCAE. The results mainly show that:(1) the SCA base models with multiple combinations of inputs and internal parameters are able to predict the BDs with good and different performances; (2) the optimal SCA base model has three inputs, which indicates that the temperatures before breakup and just after freeze-up as well as the maximum of water flow in March are relatively important indicators of BD; (3) several SCA base models selected with two criteria, i.e. the best validation performance (criterion 1) versus better training and validation performance (criterion 2), are searched for further combination. The models using criterion 2 tend to have more cutting and merging operations during the training process than those with criterion 1; (4) most of the ensemble models based on the SAM improve upon the base models in terms of both training and validation performance; The optimal stacking ensemble model ($enSAM_5$) has the lowest $RMSE_{avg}$, which improves upon the optimal SCA base model SCA_7 by 25.3 %, (5) the $enSAM_5$ is the average output from four SCA base models, where one base model belongs to criterion 1 and the other three base model belong to criteria 2. It indicates the different model selection criteria do improve the diversity and thus further help to improve the performance of ensemble models. This first application of the SCAE to river ice forecasting highlights the possibility of using the ensemble learning paradigm to enhance the SCA's predictive performance. The potential applications of the SCAE to other forecasting problems are expected.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This research was supported by the National Key Research and Development Plan (2016YFA0601502), National Natural Science Foundation of China (61601522), and the Hundred Talents Program of Sun YatSen University. The first author would like to acknowledge the support of his previous colleagues at River Forecast Center of Alberta Environment and Parks: Bernard Trevor, Nadia Kovachis Watson, Stefan Emmer, Patricia Stevenson, Evan Friesenhan, Andun Jevne and Chandra Mahabir.

References

Anctil, F. and Lauzon, N. (2004). Generalisation for neural networks through data sampling and training procedures, with applications to streamflow predictions. *Hydrology and Earth System Sciences*, 8(5), 940-958. https://doi.org/10.5194/hess-8-940-2004

Andrishak, R. and Hicks, F. (2011). Ice effects on flow distributions within the Athabasca Delta, Canada. *River Research and Applications*, 27(9), 1149-1158. https://doi.org/10.1002/rra.1414

Beltaos, S. (2000). Advances in river ice hydrology. Hydrological Processes, 14(9), 1613-1625. https://doi.org/10.1002/1099-1085(20000630)14:9<1613:AID-HYP73>3.0.CO;2-V

Beltaos, S. and Burrell, B. (2015). Hydrotechnical advances in Canadian river ice science and engineering during the past 35 years Canadian. *Journal of Civil Engineering*, 42(9), 583-591. https://doi.org/10.1139/cjce-2014-0540

Beltaos, S. and Prowse, T. (2009). River-ice hydrology in a shrin- king cryosphere. *Hydrological Processes*, 23(1), 122-144. https://doi.org/10.1002/hyp.7165

Bieniek, P.A. and Bhatt, U.S. (2011). Large-scale climate con- trols of interior Alaska river ice breakup. *Journal of Climate*, 24(1), 286-297. https://doi.org/10.1175/20 10JCLI3809.1

Breiman, L., Friedman, J., Stone, C.J., and Olshen, R.A. (1984). *Classification and regression trees*. CRC press.

Cheng, G., Dong, C., Huang, G., Baetz, B.W., and Han, J. (2016). Discrete principal-monotonicity inference for hydro-system analysis under irregular nonlinearities, data uncertainties, and multivariate dependencies. Part I: methodology development. *Hydrological Processes*, 30(23), 4255-4272. https://doi.org/10.1002/hyp.

10909

Cooley, S.W. and Pavelsky, T.M. (2016). Spatial and temporal patterns in Arctic river ice breakup revealed by automated ice detection from MODIS imagery. *Remote Sensing of Environment*, 175, 310-322. https://doi.org/10.1016/j.rse.2016.01.004

De'ath, G. and Fabricius, K.E. (2000). Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology*, 81(11), 3178-3192. https://doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2

Dietterich, T.G. (2000). Ensemble methods in machine learning. In: Kittler, J., Roli, F., Multiple Classifier Systems. *Lecture Notes in Computer Science*, pp. 1-15. https://doi.org/10.1007/3-540-45014-9_1

Erdal, H.I. and Karakurt, O. (2013). Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms. *Journal of Hydrology*, 477, 119-128. https://doi.org/10.1016/j.jhydrol.2012.11.015

Fan, Y.R., Huang, G.H., Li, Y.P., Wang, X.Q., Li, Z., and Jin, L. (2017). Development of PCA-based cluster quantile regression (PCA-CQR) framework for streamflow prediction: Application to the Xiangxi River Watershed, China. *Applied Soft Computing*, 51, 280-293. https://doi.org/10.1016/j.asoc.2016.11.039

Fan, Y.R., Huang, W., Huang, G.H., Li, Z., Li, Y.P., Wang, X.Q., Cheng, G.H., and Jin, L. (2015). A stepwise-cluster forecasting approach for monthly streamflows based on climate teleconnections. *Stochastic Environmental Research and Risk Assessment*, 29(6), 1557-1569. https://doi.org/10.1007/s00477-015-1048-y

Galelli, S. and Castelletti, A. (2013). Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling. *Hydrology and Earth System Sciences*, 17(7), 2669-2684. https://doi.org/10.5194/hess-17-2669-2013

Gao, G., Yu, G., Wang, Z., and Li, S. (2012). Advances in Break-up Date Forecasting Model Research in the Ningxia- Inner Mongolia Reach of the Yellow River, *21st IAHR International Symposium on Ice*. Dalian University of Technology Press, Dalian, China.

Guo, Q.Z. (2002). Applicability of criterion for onset of river ice breakup. *Journal of Hydraulic Engineering-Asce*, 128(11), 1023-1026. https://doi.org/10.1061/(ASCE)0733-9429(2002)128:11(1023)

Han, J.C., Huang, Y.F., Li, Z., Zhao, C.H., Cheng, G.H., and Huang. P.F. (2016). Groundwater level prediction using a SOM-aided stepwise cluster inference model. *Journal of Environmental Management*, 182, 308-321. https://doi.org/10.1016/j.jenvman.2016.07.069

He, L., Huang, G.H., and Lu, H.W. (2008a). Health-risk-based groundwater remediation system optimization through clusterwise linear regression. *Environmental Science & Technology*, 42(24), 9237-9243. https://doi.org/10.1021/es800834x

He, L., Huang, G.H., Lu, H.W., and Zeng, G.M. (2008b). Optimization of surfactant-enhanced aquifer remediation for a laboratory BTEX system under parameter uncertainty. *Environmental Science & Technology*, 42(6), 2009-2014. https://doi.org /10.1021/es071106y

Hicks, F. (2009). An overview of river ice problems: CRIPE07 guest editorial. *Cold Regions Science and Technology*, 55(2), 175-185. https://doi.org/10.1016/j.coldregions.2008.09.006

Huang, G.H. (1992). A stepwise cluster-analysis method for predicting air-quality in an urban-environment. *Atmospheric Environment Part B-Urban Atmosphere*, 26(3), 349-357. https://doi.org/10.1016/0957-1272(92)90010-P

Huang, G.H., Huang, Y.F., Wang, G.Q., and Xiao, H.N. (2006). Development of a forecasting system for supporting reme- diation design and process control based on NAPL - biodegradation simulation and stepwise-cluster analysis. *Water Resources Research*, 42(6). https://doi.org/10.1029/2005WR004006

Kowalczyk, T. and Hicks, F. (2003). Observations of dynamic ice jam release on the Athabasca River at Fort McMurray, AB. Proc. *12th Workshop on River Ice*, Edmonton, June 19-20.

Li, Z., Huang, G.H., Han, J.C., and Wang, X.Q. (2015). Development of a Stepwise-Clustered Hydrological Inference Model. *Journal of Hydrologic Engineering*, 20(10). https://doi.org/10.1061/(ASCE)HE.1943-5584.0001165

Mahabir, C., Hicks, F., and Fayek, A.R. (2006). Neuro-fuzzy river ice breakup forecasting system. *Cold Regions Science and Technology*, 46(2), 100-112. https://doi.org/10.1016/j.cold regions.2006.08.009

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine*, 6(3), 21-44. https://doi.org/10.1109/MCAS.2006.1688199

Qin, X.S., Huang, G.H., and Chakma, A. (2007). A stepwise-inference-based optimization system for supporting remediation of petroleum-contaminated sites. *Water Air and Soil Pollution*, 185(1-4), 349-368. https://doi.org/10.1007/s11270-007-9458-1

Quinlan, J.R. (1992). Learning with continuous classes, *5th Australian joint conference on artificial intelligence*. Singapore, pp. 343-348.

Rao, C.R. (1952). *Advanced Statistical Methods in Biometric Research*, Wiley, New York, pp. 106-207.

Rao, C.R. (1965). Linear Statistical Inference and Its Applications. Wiley, New York, pp. 239-301.

She, Y.T., Andrishak, R., Hicks, F., Morse, B., Stander, E., Krath, C., Keller, D., Abarca, N., Nolin, S., Tanekou, F.N., and Mahabir, C. (2009). Athabasca River ice jam formation and release events in 2006 and 2007. *Cold Regions Science and Technology*, 55(2), 249-261. https://doi.org/10.1016/j.coldregions.2008.02.004

Sun, W. (2018). River ice breakup timing prediction through stacking multi-type model trees. *Science of the Total Environment*, 644: 1190-1200. https://doi.org/10.1016/j.scitotenv.2018.07.001

Sun, W., Huang, G.H., Zeng, G.M., Qin, X.S., and Sun, X.L. (2009). A stepwise-cluster microbial biomass inference model in food waste composting. *Waste Management*, 29(12), 2956-2968. https://doi.org/10.1016/j.wasman.2009.06.023

Sun, W., Huang, G.H., Zeng, G.M., Qin, X.S., and Yu, H. (2011). Quantitative effects of composting state variables on C/N ratio through GA-aided multivariate analysis. *Science of The Total Environment*, 409(7), 1243-1254. https://doi.org/10.1016/j.scitotenv.2010.12.023

Sun, W. and Trevor, B. (2017). Combining k-nearest-neighbor models for annual peak breakup flow forecasting. *Cold Regions Science and Technology*, 143, 59-69. https://doi.org/10.1016/j.coldregions.2017.08.009

Sun, W. and Trevor, B. (2018a). Multiple Model Combination Methods for Annual Maximum Water Level Prediction during River Ice Breakup. *Hydrological Processes*. https://doi.org/10.1002/hyp.11429

Sun, W. and Trevor, B. (2018b). A stacking ensemble learning framework for annual river ice breakup dates. *Journal of Hydrology*, 561, 636-650. https://doi.org/10.1016/j.jhydrol.2018.04.008

Sun, W., Trevor, B., and Kovachis, N. (2015). *Athabasca River Ice Observations 2014-2015 (Annual Report)*, Alberta Environment and Parks Edmonton, Alberta.

Tatsuoka, M.M. (1971). *Multivariate Analysis*. Wiley, New York, pp. 38-197.

Ting, K.M. and Witten, I.H. (1999). Issues in stacked generalization. *Journal of Artificial Intelligence Research*, 10, 271-289. https://doi.org/10.1613/jair.594

Wang, G., Hao, J., Ma, J., and Jiang, H. (2011). A comparative assessment of ensemble learning for credit scoring. *Expert Systems with Applications*, 38(1), 223-230. https://doi.org/10.1016/j.eswa.2010.06.048

Wang, S., Huang, G.H., and He, L. (2012). Development of a cluster-wise-linear-regression-based forecasting system for characterizing DNAPL dissolution behaviors in porous media. *Science of the Total Environment*, 433, 141-150. https://doi.org/10.1016/j.scitotenv.2012.06.045

Wang, T., Yang, K.L., and Guo, Y.X. (2008). Application of artificial

neural networks to forecasting ice conditions of the Yellow River in the Inner Mongolia reach. *Journal of Hydrologic Engineering*, 13(9), 811-816. https://doi.org/10.1061/(ASCE)1084-0699(2008)13:9(811)

Wang, X.Q., Huang, G.H., Lin, Q.G., Nie, X.H., Cheng, G.H., Fan, Y.R., Li, Z., Yao, Y., and Suo, M.Q. (2013). A stepwise cluster analysis approach for downscaled climate projection - a Canadian case study. *Environmental Modelling & Software*, 49, 141-151. https://doi.org/10.1016/j.envsoft.2013.08.006

Wang, X.Q., Huang, G.H., Zhao, S., and Guo, J.H. (2015). An open-source software package for multivariate modeling and clustering: applications to air quality management. *Environmental Science and Pollution Research*, 22(18), 14220-14233. https://doi.org/10.1007/s11356-015-4664-7

Wang, Y. and Witten, I.H. (1997). Inducing model trees for continuous classes, *Proceedings of the Ninth European Conference on Machine Learning*, pp. 128-137.

Wilks, S.S. (1962). *Mathematical Statistics*. Wiley, New York, pp. 20-209.

Wolpert, D.H. (1992). Stacked generalization. *Neural Networks*, 5(2), 241-59. https://doi.org/10.1016/S0893-6080(05)80023-1

Zaier, I., Shu, C., Ouarda, T.B.M.J., Seidou, O., and Chebana, F. (2010). Estimation of ice thickness on lakes using artificial neural network ensembles. *Journal of Hydrology*, 383(3-4), 330-340. https://doi.org/10.1016/j.jhydrol.2010.01.006

Zhai, Y., Huang, G.H., Wang, X.Q., Zhou, X., Lu, C., and Li, Z. (2019). Future projections of temperature changes in Ottawa, Canada through stepwise clustered downscaling of multiple GCMs under RCPs. *Climate Dynamics*. 52(5), 3455-3470. https://doi.org/10.1007/s00382-018-4340-y

Zhao, L., Hicks, F., Fayek, A.R., and Kovachis, N. (2010). Forecasting the Onset of Breakup using Artificial Neural Networks, *20th IAHR International Symposium on Ice*.

Zhao, L., Hicks, F.E., and Robinson Fayek, A. (2012). Applicability of multilayer feed-forward neural networks to model the onset of river breakup. *Cold Regions Science and Technology*, (70), 32-42. https://doi.org/10.1016/j.coldregions.2011.08.011

Zhao, S., Huang, G.H., Cheng, G.H., Sun, W., Su, Q., Tao, Z.Y., and Wang, S.G. (2017). A Stepwise-Cluster Inference Model for Phenanthrene Immobilization at the Aqueous/Modified Palygorskite Interface. *Water*, 9(8). https://doi.org/10.3390/w9080590

Zhou, H., Li, W., Zhang, C., and Liu, J. (2009). Ice breakup forecast in the reach of the Yellow River: the support vector machines approach. *Hydrology and Earth System Sciences Discussions*, 6(2), 3175-3198. https://doi.org/10.5194/hessd-6-3175-2009

Zhuang, X.W., Li, Y.P., Huang, G.H., and Liu, J. (2016a). Assessment of climate change impacts on watershed in cold-arid region: an integrated multi-GCM-based stochastic weather generator and stepwise cluster analysis method. *Climate Dynamics*, 47(1-2), 191-209. https://doi.org/10.1007/s00382-015-2831-7

Zhuang, X.W., Li, Y.P., Huang, G.H., and Wang, X.Q. (2016b). A hybrid factorial stepwise-cluster analysis method for streamflow simulation - a case study in northwestern China. *Hydrological Sciences Journal*, 61(15), 2775-2788. https://doi.org/10.1080/02626667.2015.1125482