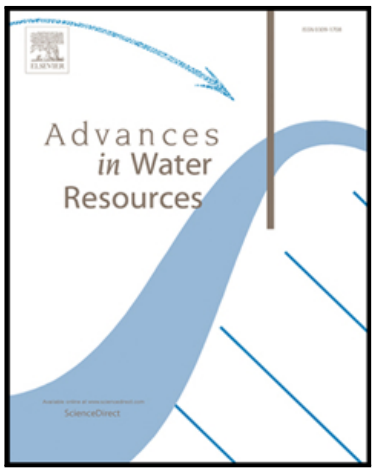

Diversity-driven ANN-based ensemble framework for seasonal low-flow analysis at ungauged sites

Mohammad H. Alobaidi , Taha B.M.J. Ouarda ,
Prashanth R. Marpu , Fateh Chebana

PII: S0309-1708(20)30180-9
DOI: https://doi.org/10.1016/j.advwatres.2020.103814
Reference: ADWR 103814

To appear in: *Advances in Water Resources*

Received date: 24 February 2020
Revised date: 19 October 2020
Accepted date: 10 November 2020

Please cite this article as: Mohammad H. Alobaidi , Taha B.M.J. Ouarda , Prashanth R. Marpu , Fateh Chebana , Diversity-driven ANN-based ensemble framework for seasonal low-flow analysis at ungauged sites, *Advances in Water Resources* (2020), doi: https://doi.org/10.1016/j.advwatres.2020.103814

Highlights

- A novel ensemble-based machine learning framework is proposed to estimate seasonal low-flow at ungauged sites.
- The concept of information mixture is utilized in the ensemble training and ensemble integration stages.
- Regressive sub-model integration techniques are used in the combining stage to create robust ensemble forecasts.
- The model provided improved performance, compared to other models, when applied to a case study in Canada.

Diversity-driven ANN-based ensemble framework for seasonal low-flow analysis at ungauged sites

Mohammad H. Alobaidi[1*], Taha B.M.J. Ouarda[2], Prashanth R. Marpu[3] and Fateh Chebana[2]

[1]Department of Civil Engineering and Applied Mechanics, McGill University, 817 Rue Sherbrooke Ouest, Montréal, Québec, Canada, H3A 0C3

[2]Eau Terre Environnement (ETE), Institut National de la Recherche Scientifique (INRS), 490 de la Couronne, Québec City, Québec, Canada, G1K 9A9

[3]Department of Electrical Engineering and Computer Science, Khalifa University, Masdar City, P.O. Box 54224, Abu Dhabi, UAE

*Corresponding author:

Email: mohammad.alobaidi@mail.mcgill.ca

Address:

Department of Civil Engineering and Applied Mechanics, McGill University, 817 Rue Sherbrooke Ouest, Montréal, Québec, Canada, H3A 0C3

Abstract

Low-flow estimation at ungagged sites is a challenging task. Ensemble-based machine learning regression has recently been utilized in modeling hydrologic phenomena and showed improved performance compared to classical regional regression approaches. Ensemble modeling mainly revolves around developing a proper training framework of the individual learners and combiners. An ensemble framework is proposed in this study to drive the generalization ability of the sub-ensemble models and the ensemble combiners. Information mixtures between the subsamples are introduced and, unlike common ensemble frameworks, are explicitly devoted to the ensemble members as well as ensemble combiners. The homogeneity paradigm is developed via a two-stage resampling approach, which creates sub-samples with controlled information mixture levels for the training of the individual learners. Artificial neural networks are used as sub-ensemble members in combination with a number of ensemble integration techniques. The proposed model is applied to estimate summer and winter low-flow quantiles for catchments in the province of Québec, Canada. The results provide significant improvement when compared to the other models presented in the literature. The results of the homogeneity levels from the optimum ensemble models demonstrate the importance of utilizing the diversity concept in ensemble learning applications.

Keywords: Ensemble Learning; Information Theory; Diversity-in-Learning; Low-Flow Estimation.

1. Introduction

Reliable low-flow estimates are important for a large number of engineering applications such as water quantity and quality management, and environmental impact assessment. Low-flow quantile estimates can be obtained using a number of approaches, such as flow duration curves or best-fit probabilistic distribution. Both approaches require the availability of low-flow information at the site of interest. When streamflow data is not available (ungauged sites), regional techniques are used to estimate the low-flow statistics. Low-flow estimation techniques at ungauged sites include regional prediction curves, spatial interpolation and regional mapping, synthetic streamflow time series for low-flow estimation, and regional regression modeling (Smakhtin, 2001b). Low-flow estimation is well-established in the literature and detailed information can be found in (Gustard and Demuth, 2009, Ouarda et al., 2008a, Smakhtin, 2001a). Among the various methods for low-flow estimation at ungauged sites, regional regression techniques are commonly used in practice for low-flow estimation at ungauged sites (Vogel and Kroll, 1990, Vogel and Kroll, 1992, Dingman and Lawlor, 1995, Ouarda and Shu, 2009). A classical regression technique (Thomas and Benson, 1970) for such task has the following generalized form:

$$Q_{d,T} = \alpha \prod_{i=1}^{l} x_i^{\beta_i}, \qquad (1)$$

where $Q_{d,T}$ is the T-year low-flow quantile corresponding to a duration of d-days at the site of interest; x_i is the i^{th} variable (site characteristic) used for low-flow quantile estimation; β_i is the i^{th} model parameter which needs to be estimated; l is the total number of site characteristics used in the model and α is the multiplicative error term.

Logarithmic transformation linearizes the model governed by Equation (1). A multiple linear regression (MLR) is then used to estimate it. A disadvantage of this model is that logarithmic transformation may result in a bias in the estimation of its parameters (Kouider, 2003, Ouarda et al., 2008b, Shu and Ouarda, 2008, McCuen et al., 1990). In case of low-flow estimation, such bias may result in significantly improper model performance (Ouarda and Shu, 2009).

Ensemble modeling for regression applications can tackle the different challenges manifesting in low-flow estimation at ungauged sites. In fact, this study is aimed at building on Ouarda and Shu (2009). An ensemble framework is presented in the present paper, where the architecture of its three phases (resampling, training and combining) targets implicitly and empirically optimized generalization ability. Diversity-controlled approach is embedded in a proposed multi-stage resampling approach. The ensemble members and the ensemble combiner are sub-sequentially optimized for enhanced ensemble estimation performance. In this article, specifics of the proposed approach are presented. The results of the proposed model are compared with the results from the previous models on the same case study. The proposed ensemble framework for the problem of low-flow estimation at ungagged sites is intended to show how the physical nature of the problem of interest inspires the design of the ensemble architecture for improved generalization ability. The main contributions of the present work are listed as follows:

- The research work presents a generalized ensemble model which has been inspired from the concept of diversity-in-learning and the problem of interest.
- The ensemble framework requires relatively reduced computing resource to be trained and validated in a reasonable timeframe.

- The ensemble framework is capable of parallelized training routine for efficient allocation of computing resource.

- The model is theoretically scalable in its own size, the available features as well as the available observations. The ensemble framework is also of parallel learning nature, allowing for efficient computational routine.

The structure of the paper is as follows; in Section 2, ensemble learning with artificial neural network sub-models is discussed. In Section 3, the proposed ensemble approach is provided. In Section 4, a detailed description of the case study is presented. Section 5 describes the experimental setup, model-specific configurations, for the present work. In Section 6, the study results are discussed. Lastly, Section 7 summarizes the study conclusions and provides recommendations for future research work.

2. Background

2.1. Brief overview of ensemble learning

Ensemble regression modeling is an evolving field in machine learning, which allows remedy to the nature (feature space) and availability (sample size) challenges of the data. In regard to the present application, low-flow estimation at ungauged sites utilizes a relatively limited number of covariates, disqualifying the use of deep learning models. Moreover, shallow machine learning models inherently suffer from instability challenges (training leads to different local minima of the parameter choices), which are exacerbated in the case of limited training data. Ensemble Learning provides a solution to these two major issues.

An ensemble model generally comprises a set of regression models (known as sub-ensembles, individual learners, ensemble members or predictors). Ensemble learning defines the technique upon which the information from the dataset is distributed to the sub-ensembles, for training, as well as the combination plan of the sub-ensembles estimates toward an observation (Dong et al., 2020). Many research efforts in the literature have provided empirical and theoretical evidence toward ensemble models' superiority in performance and generalization ability (Chen et al., 2012, Dietterich, 2000, Green and Ohlsson, 2007, Hansen and Salamon, 1990, Maclin and Opitz, 1999, Mendes-Moreira et al., 2012, Vrugt and Robinson, 2007, Zhang and Ma, 2012).

Ensemble modeling can be divided into three main stages; resampling, generation and training, and integration. In the resampling phase, the dataset, or sample, undergoes a pre-defined process which ultimately creates the sub-samples, utilized for training the individual members. Several resampling plans exist in the literature such as the different bootstrap resampling techniques (Efron, 1982, Bühlmann, 2003). In the ensemble model generation and training second phase, the sub-ensembles are created and arranged to learn the functional relationship between the explanatory and response variables, using the available information from the sub-samples. The sub-ensembles can be any regression model which seen best for the system of interest. A homogenous ensemble composes of similar sub-ensemble models (the same model structure and unknown parameters to be solved). In this case, the variation in the information by the different sub-samples will prompt diverse solutions in the sub-ensemble's parameters. In the case of nonhomogeneous ensembles, the sub-ensembles can be a collection of different regression models. An ensemble of the same individual members can still be nonhomogeneous if they comprise different sub-ensemble topologies, such as ANNs with different configuration. A

popular example of nonhomogeneous ensembles is Random Forest (RF). This model utilizes a multitude of classification and regression trees (CARTs), generated via random subspace resampling, where each CART is expected to train over a number of the available feature space (Ho, 1995, Ho, 1998).

The diversification in the resamples (sub-samples) and the sub-ensembles will produce even more diverse relationship in the nonhomogeneous ensemble models (Zhang and Ma, 2012). In this case, the ensemble models, in their mathematical nature, improves the overall generalization ability of the ensemble model (Ueda and Nakano, 1996, Vrugt and Robinson, 2007). In the ensemble integration phase, a combiner is used to fuse the different estimates from the individual learners, toward one observation, into the ensemble estimate. The choice of the combiners can rely on the nature of the resampling techniques and the sub-ensembles chosen for the ensemble model. Generally, a combiner can be as simple as taking the mean of the individual learners' estimates. A combiner can also be as complicated as a final-regression model on the sub-ensembles' estimates. Such combiner is usually tuned in the training stage of the ensemble, using the complete training set, the pre-defined training subsamples, or different training plans.

Examples of popular ensemble models are Bagging (Breiman, 1996a), Stacking (Breiman, 1996b, Wolpert, 1992) and Boosting (Bühlmann and Hothorn, 2007, Drucker, 1997, Duffy and Helmbold, 2002, Freund and Schapire, 1996, Friedman et al., 2000, Friedman, 2001, Sharkey, 1999). Bagging (also known as Bootstrap Aggregating) utilizes bootstrap resampling to generate the sub-samples which are used to train the sub-ensembles, while the combiner in this model are simply the arithmetic mean of the sub-ensemble estimates. It is worth noting that diversity generating mechanism in RFs is also an extension of Bagging (Breiman, 2001). In stacking, the creation of the sub-samples can be provided using a resampling plan. Once the sub-ensembles are

trained using their individual sub-samples, a linear combiner of the sub-ensemble estimates is trained. Non-negative weights are computed for combining the sub-ensemble estimates into an ensemble estimate. These two ensemble models have the advantage of relatively fast training, as the re-samples and the sub-ensemble are created in parallel.

On the other hand, the combiner used in Boosting requires an in-series creation of the sub-samples as well as the ensemble members; this ensemble plan starts by training one sub-ensemble using all the available information in the training set. The estimation error associated with each training instance is computed and compared. The second sub-sample is a sampling with replacement from the original sample set. Further, the instances with high estimation error will have a greater probability of being selected in the second sub-sample in order to focus the training of the second predictor on such instances. This process of sub-sample creation and predictor training is carried out until a stopping criterion is satisfied. The trained predictors will then be provided with combination weights (proportional to their accuracy) that combine the sub-estimates into an ensemble estimate. This ensemble model can be slow and highly sensitive to outliers.

Other techniques are used in the final stage of ensemble learning; one notable approach is Bayesian Model Averaging (BMA). This approach is suggested by Learner (1978) and recently proliferated in the applied field (Duan et al., 2007, Dong et al., 2013, Qu et al., 2017, Huo et al., 2019). Contrary to the name, BMA is actually a selection method. BMA does not combine sub-ensemble inferences, but rather selects the sub-ensemble to which the target observation supposedly belongs. As such, each sub-ensemble is considered as a Data-Generating Model (DGM), and for BMA to prevail, one of the DGMs should be the true model. Under the givens of the present work, it is not reasonable that one of the trained models will be the true DGM to any

of the low-flow quantiles. Hence, a combination of the available inference (which is the motivation behind ensemble learning) is expected to produce better generalization ability. Moreover, the explicit diversity mechanism in the proposed model is partly driven by the sub-ensemble combination rather than selection; using BMA in this framework beats the point (Clarke, 2003).

2.2. Ensemble learning in hydrology

Several studies applied ensemble learning in hydrology. For example, Francke et al. (2008) compared different methods with respect to performance in measuring the suspended sediment concentration and construction of sedigraph. This study showed that regression-based random forests and quantile random forests ensembles provided robust performance, in contrast to the inferior performance of classical linear regression approach in such problem. The study also outlined the capability of the applied ensembles in providing uncertainty assessment as well as interpretation of predictor effects. Erdal and Karakurt (2013) aimed at assessing the application of classification and regression trees (CARTs) in the bagging and boosting ensemble frameworks for streamflow forecasting. Results from a support vector regression (SVR) model were used as benchmark. The study showed that both bagging-based and boosting-based CARTs can significantly enhance the prediction accuracy when compared to a single CART model as well as the benchmark SVR model results. Further, Shu and Ouarda (2007) used bagging ensemble model for flood frequency analysis at ungauged sites. The study used the canonical correlation analysis (CCA) to draw canonical projections of the sub-ensembles' meteorological and physiographic input variables. The results indicated that the proposed CCA-based bagged ensemble has the best performance when compared to other single models. In addition, this study showed that CCA pre-processing improved the ensemble performance when compared with the

same model but using original variables space. Many studies compared the performance of different ensemble methods or different combination techniques when applied to hydrological problems (Shu and Ouarda, 2007, Ouarda and Shu, 2009, Shu and Burn, 2004, Diks and Vrugt, 2010, Vrugt and Robinson, 2007, Ajami et al., 2006). A recent review and comprehensive application of common Ensemble frameworks is presented in (Alobaidi et al., 2019).

2.3. Artificial neural networks in an ensemble framework

Artificial neural networks (ANNs) are evolving machine learning tools that can articulate the relationship between the models inputs and outputs without predefined assumptions, neither on the model parameters nor on the system variables (Bishop, 2006). ANNs have received much attention in the field of hydrology (Govindaraju and Rao, 2010). Regression-based ANNs proved to be flexible models and effective as sub-ensembles in many studies (Shu and Burn, 2004, Green and Ohlsson, 2007, Siou et al., 2011, Islam et al., 2003, Zaier et al., 2010, Agrafiotis et al., 2002, Hashem, 1993, Hashem et al., 1994). Furthermore, many studies attempted to describe the generalization ability of ANN-based ensemble models. The mathematical interpretation of the statistical performance of ensembles with ANN individuals was frequently investigated (Geman et al., 1992, Krogh and Vedelsby, 1995, Hashem, 1997, Zhou and Chen, 2002, Granitto et al., 2005, Green and Ohlsson, 2007, Alam et al., 2019). The idea behind using ANNs in an ensemble framework is to promote diversity, which can ultimately improve the generalization ability of the ensemble model beyond any of its individual members (Liu, 1999, Brown, 2004, Alam et al., 2019).

Diversity is defined as the amount of disagreement between ensemble members (Kuncheva and Whitaker, 2003). Metrics of diversity concept can be usually defined via the bias-variance-

covariance decomposition of the ensemble model (Kuncheva, 2003, Lázaro et al., 2020). Although the general concept of diversity is well defined, the research on providing clear mathematical description to diversity of an ensemble is still an open topic (Slavin Ross et al., 2019). In general, in ensemble learning, the sub-models are usually trained on resamples of the training data. For example, boosting ensembles update their sampling distribution before generating the new subset for the corresponding member of the boosting ensemble. This allows for misclassified instances, from previously generated sub-ensembles, to be selected in upcoming sub-ensembles, and the ensemble members are expected to be diverse as a result. There is no explicit measure of diversity. This diversity-manifesting mechanism is in fact native to boosting models and the formulation of diversity mechanisms drastically change among ensembles. Moreover, once such a relation is formulated, optimizing the diversity-accuracy tradeoff of the ensemble can be carried out to maximize the ensemble generalization ability (Schmidt, 2004, Brown et al., 2005a, Brown et al., 2005b, Sun and Zhou, 2018).

To this extent, ANN-based ensemble models seem to be ideal for a challenging regression problem in hydrological modeling such as regional frequency analysis at ungauged sites (Shu and Ouarda, 2007). In addition, ANN-based ensemble learning has been utilized in regional low-flow analysis. The work by Ouarda and Shu (2009) used ANN-based bootstrap aggregation ensemble, with stacking combiner, in order to provide improved summer and winter low-flow quantile estimates at ungauged sites. The ensemble approach provided improved generalization ability when compared to the single ANN model and the classical regression model. Although relatively improved, the scale challenge (significantly quantile values vary quite from one site to another) was still apparent in the ensemble model; low-flow quantiles for some of the ungauged basins were highly skewed from the general pattern and, therefore, poorly estimated.

3. Proposed Approach

The ANN framework employed in this work is ensemble-based. It is utilized to estimate the functional relationship between the explanatory variables and the target variables, inputs and outputs, respectively. Figure 1 demonstrates the detailed modeling steps for the proposed ensemble. It is worth noting that the proposed ensemble framework is a generalization of earlier work which contributed to the field of interest and detailed work on earlier versions of the proposed model can be found in (Alobaidi et al., 2015). After identifying the system's variables, and before starting the validation and the training process, pre-processing is applied on the identified inputs and outputs of the system. Pre-processing techniques range from linear transformation, such as linear scaling and normalization, to nonlinear techniques, such as logarithmic and Box-Cox transformations (Alobaidi et al., 2014). The choice of a proper pre-processing plan incorporates the type of data used, the individual members' requirements and the ensemble method itself. More about pre-processing can be found in (Ouarda et al., 2001, Ouarda et al., 2008b, Shu and Ouarda, 2008, Ouarda and Shu, 2009, Basu and Srinivas, 2014). After the pre-processed plan is determined, and the modified sample set is acquired, the proposed methodology follows systematic processes, described in the following subsections.

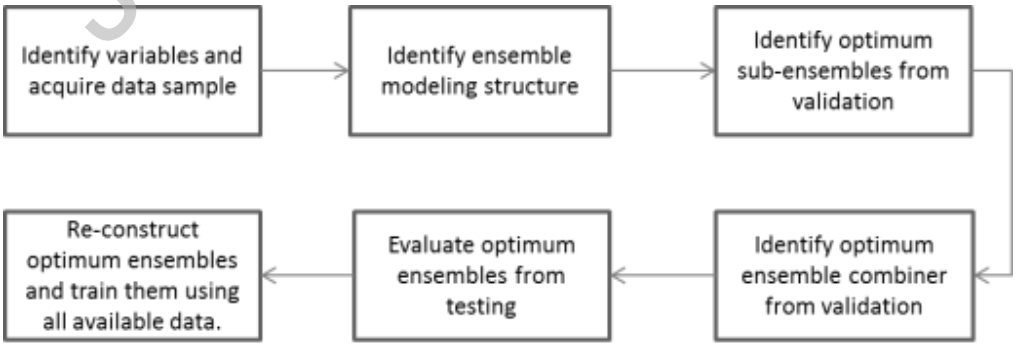

Figure 1: Modeling process of the proposed ensemble.

3.1. The resampling algorithm

A two-stage sampling process is applied in a controlled environment, where two homogeneity control (or mixture control) measures are introduced. One measure is to control the amount of information that is blocked from the members' training and used for the training of the ensemble combiner. The other parameter is introduced to promote "measured" diversity between the subsamples (or resamples). The proposed resampling technique is responsible for producing the required subsets which will be trained for sub-models. It is important to note that resamples' and sample's size annotations are used in order to track the diversity evolution with respect to the resample size. Also, this will help differentiate between the unique information and the mixed information that the first-stage and second-stage represent, respectively.

Initially, consider a sample set which corresponds to size *N* available for the training process, a size-controlled part of the training data will be chosen randomly (sampling without replacement) and blocked from the related training sample as follows:

$$n_{blocked} = N \times m_c, \quad (2)$$

where $n_{blocked}$ refers to the size of the relevant blocked observations and m_c refers to the mixture ratio that is calculated to measure a number typically between 0% and 30% which refers to the percentage of blocked information. The percentage often depends on the limited access for training data, the size of the ensemble, as well as the type of the ensemble combiner.

The subsample size of the first-stage is computes as follows:

$$n_1 = \frac{(N - n_{blocked})}{S}, \qquad 1 \leq S \leq (N - n_{blocked}), \qquad (3)$$

where n_1 refers to the size of the subsample, and *S* refers to the size of the ensemble. Every subsample has different observations even though subsamples have the same size. In other words, an observation cannot be found in multiple subsamples.

Successively, the subsample size of the second-stage is computed by observing the amount of exchanged information between subsamples obtained from the first stage, as shown in Equation (4). This is done after defining a specific parameter which is going to control the information mixture in each subsample. Further, the mixture-control parameter can be defined as a function or set of functions that can be given to each first-stage resample (or each individual member's training set). By doing so, the nature of the relationship between each second-stage resample can be different and reshaped in a more flexible way, if required. The second-stage subsample size is then computed as:

$$n_{2_i} = n_{1_i} + \left[\sum_{\substack{j=1 \\ j \neq i}}^{S} \left(f(m_e)_{ij} \times n_{1_j} \right) \right], i, j = 1,2,3,..,S, \qquad (4)$$

where n_{2_i} refers to the size of the i^{th} second-stage subsample, and $f(m_e)_{ij}$ refers to the mixture ratio or the relationship between the j^{th} first-stage subsample and the i^{th} second-stage subsample. n_{1_j} refers to the size of the j^{th} first-stage subsample, while n_{1_i} refers to the size of the i^{th} first-stage subsample. Note that the subscript is not removed to indicate that first-stage resamples have unique information.

$f(m_e)_{ij}$, which are the mixture ratios, can be designed to take different mathematical forms, such as a linear relation or a quadratic function. In all the relationships, this parameter is between 0 and 1; where zero indicates that the resamples from the both stages are exactly similar. In other words, there is no mixture. Saturation, on the other hand, indicates that the resamples from the second-stage process are a duplicate of the observations available in the original sample. These two "extremes" (0 and 1) are avoided in the mixture parameters, because the zero value may downgrade the diversity for the use of first-stage resamples, and the saturation of the resamples is a result of the individual member models instead of the training resamples. Moreover, one can observe that the individual diversity parameters are between 0 and 1, and they may (or may not) sum to 1, depending on the ensemble size and the individual value of the parameter. Hence, Equation 4 is not a weighted sum. Two different relationships of the mixture parameters are defined in Table 1. Additionally, Figure 2 illustrates the effect of the chosen mixture relationship on the amount of information dedicated for each ensemble member, relative to the original sample after blocking the random subsample used for training the ensemble combiner. The final mixture measure can take many forms which are mapped from the mixture parameters. Equation 7 shows how they can be computed for the present work. Also, the choice of a link function is arbitrary. This is analogues to the choice of a transfer function for the ANN model's hidden neurons, or a training algorithm. An empirical evaluation usually presents the best link function for a given case study.

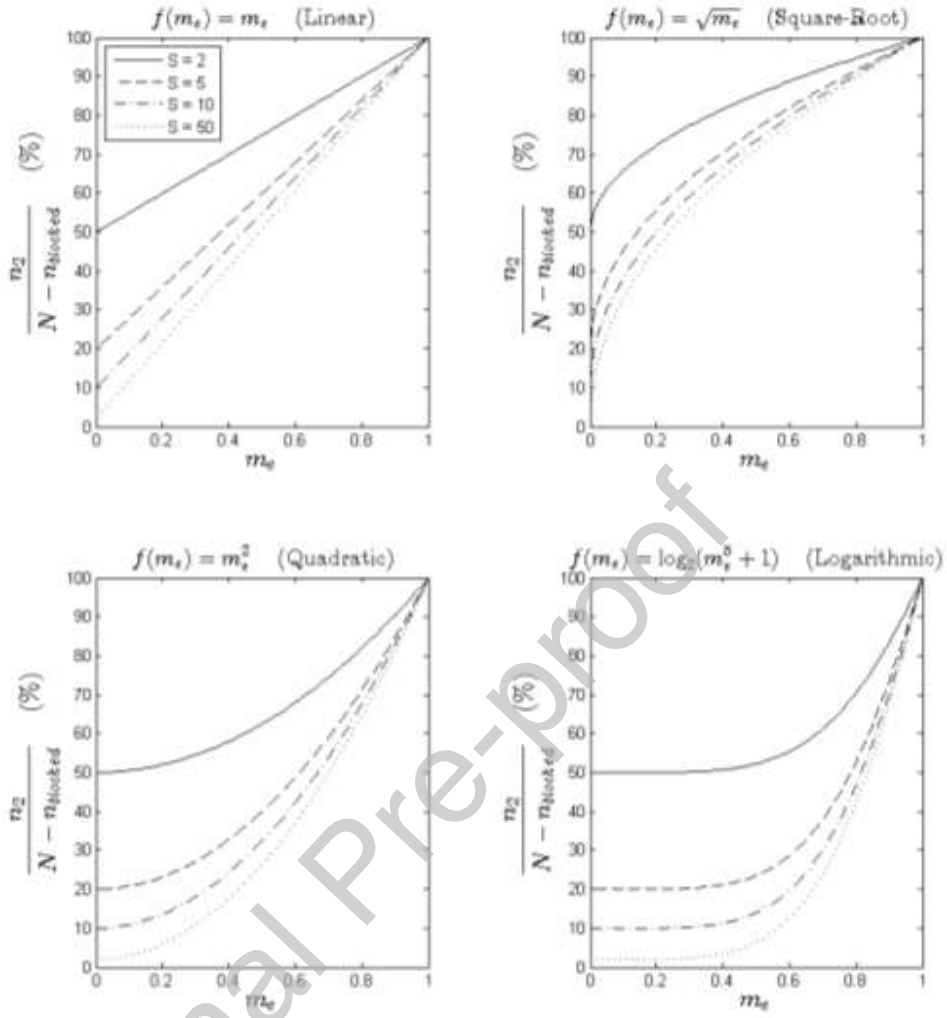

Figure 2: Graphical representation of different mixture relations.

Table 1: Different mathematical forms of the mixture parameters.

Relationship	Mathematical Representation	
Linear	$f(m_e)_{ij} = a_{ij} \times m_{e_{ij}} + b_{ij}$	(5)
Power	$f(m_e)_{ij} = a_{ij} \times \left(m_{e_{ij}}\right)^{c} + b_{ij}$	(6)

In this work, a linear mixture ratio, or homogeneity ratio, definition is used. As a consequence, the degree of information mixture for the second-stage resamples is computed as follows:

$$f(m_e)_j = m_{e_j} = \frac{n_{shared_j}}{n_{1_j}}, \qquad 0 \leq n_{shared_j} \leq n_{1_j}, \tag{7}$$

where n_{shared_j} refers to the number of observations related to the j^{th} subsample, and n_{1_j} refers to the size of the j^{th} first-stage subsample. Note that the constraints in equation (5) limit the j^{th} homogeneity ratio, m_{e_j}, by restricting its value to be between zero and one. This means that the amount of obtained information in the j^{th} first-stage subsample should not go above its size in order to prevent redundancy in shared information. This constraint, as a consequence, is what makes the proposed resampling technique very different from the conventional concept of bagging. m_{e_j}, which is the mixture ratio, is set a priory and, by rearranging equation (5), the size of the exchanged information, n_{shared_j}, is now calculated. Furthermore, the relation described in equation (5), specifically, and equation (4), generally, can be further generalized to have different

information mixture ratios. Such approach allows for the mixture ratio relation, $f(m_e)$, to have conditional information sharing, such that:

$$M = \begin{bmatrix} 0 & f(m_e)_{1,2} & f(m_e)_{1,3} & \dots & f(m_e)_{1,S} \\ f(m_e)_{2,1} & 0 & f(m_e)_{2,3} & \dots & f(m_e)_{2,S} \\ f(m_e)_{3,1} & f(m_e)_{3,2} & 0 & \dots & f(m_e)_{3,S} \\ \vdots & \vdots & \vdots & f(m_e)_{i,j} & \vdots \\ f(m_e)_{S,1} & f(m_e)_{S,2} & f(m_e)_{S,3} & \dots & 0 \end{bmatrix}, \quad (8)$$

where M refers to the generalized-global mixture ratios' matrix, $f(m_e)_{i,j}$ is the mixture ratio relation which controls the information exchanged by the i^{th} first-stage subsample, dedicated to the j^{th} first-stage subsample, making-up the j^{th} second-stage subsample. M is set to be symmetric. Hence, equation (5) defines the identical amount of mixture shared, either way, by two first-stage resamples.

Furthermore, setting the mixture ratios to be equal results in having the amount of shared information to be the same for every first-stage subsample, n_{shared_j}, such that:

$$n_{shared_1} = n_{shared_2} = \cdots = n_{shared_S} = n_{shared}, \quad (9)$$

or:

$$m_{e_1} = m_{e_2} = \cdots = m_{e_S} = m_e. \quad (10)$$

The last constraint reduces equation (4) to the following expression:

$$n_2 = \frac{N \times (1 - m_c)}{S} \times \left((m_e \times S) - m_e + 1\right) \quad (11)$$

or:

$$n_2 = N \times \left(\frac{(1 - m_c) \times m_e}{S}\right) \times \left(\frac{1 + (m_e \times S)}{m_e} - 1\right). \qquad (12)$$

In a computational framework, the indices of one variable of the system input and output variables can be used, on which the resampling algorithm can be done. Then, the subsets, or resamples, are obtained by retrieving the observations relating to the indices as well as the corresponding observations of the other variables. After calculating the size of the corresponding ensemble model, *S*, a study of various homogeneity ratios is conducted in the ensemble validation process in order to arrive at the ideal estimation of unique information to be shared by the first-stage subsamples.

The reader should note that the proposed resampling approach calculates the amount of information mixture in the subsamples; however, we execute a random selection of the amount of shared information by determining the mixture ratio, *m*. It is expected to yield improved generalization results by providing the sub-ensemble models just-enough information about the connection between the explanatory and the specified target variables and then combining the inferences from the individual learners.

3.2. Sub-ensemble model

The MLP-based ANN ensemble model (ANN-E) is used. The ANN sub-models have only one input and output layer. Also, they only have one hidden layer. Moreover, the number of explanatory variables is determined based on the number of neurons in the input layer as they are set equal. Similarly, the number of response variables depends on the number of neurons in the output layer as they are set equal. Determining the number of hidden neurons is performed in the validation procedure. For each hidden neuron, the transfer function (or activation function) is

selected as the tan-sigmoid. Also, the output layer employs a linear transfer function for its neuron. This output layer configuration is common in regression based ANNs.

The training algorithm used is the Levenberg-Marquardt (LM) algorithm (Hagan and Menhaj, 1994) which outperforms the gradient descent approach (Hagan and Menhaj, 1994, Shu and Ouarda, 2007). μ is adopted as a scalar parameter while working with the LM algorithm. A relatively large scalar prompts the algorithm to stipulate the gradient descent method. On the other hand, a lower-magnitude scalar drives the algorithm to stipulate the Gauss-Newton method (Demuth et al., 2006). Such method is considered more accurate in obtaining a global optimum. It is important to circumvent the over-fitting problem in trained ANN-E models by regularization and specifying stopping criteria (Bishop, 2006, Tikhonov and Arsenin, 1979, Vapnik, 1998). In this work, an early stopping criterion is specified in the training process (Tetko and Villa, 1997, Hagiwara, 2002, Bühlmann and Hothorn, 2007). After that, we introduce the validation procedure which is demonstrated in the next section.

Finding the optimum solution for the ensemble size, *S*, is often computationally expensive. The ensemble size, *S*, is primarily responsible for defining the size of the first-stage resamples. It is also used to find the size of the shared information, besides the mixture ratios, m_e and m_c, in the second-stage resamples. One way to reduce computational cost, when the utilized dataset is large, is through changing and assigning a reasonable value for the ensemble size, given the available training information. However, when utilizing a small or limited dataset, further due diligence is required by investigating the optimum ensemble size. This is done by validating and comparing different ensemble models. As a result, the ensemble size will produce sufficient training observations in order to perform the training of the ensemble members successfully.

Also, it will achieve generalization over the entire space of the target variable. In the proposed study, different values of the ensemble size, *S*, are investigated along with the mixture ratio.

3.3. Techniques for ensemble integration

The final stage in ensemble learning is the ensemble integration. Estimates from the individua members are integrated into one ensemble estimate. In this study, four ensemble integration techniques are utilized; the mean, median, OLS linear regression and linear robust fitting. It is recommended to use the mean statistic in order to examine the normality of the distribution of the estimates. Similarly, Bagging suggests the choice of the mean statistic as an ensemble integration technique (Breiman, 1996a). The ensemble estimate of the i^{th} observation, $\hat{y}_{e,i}$, is obtained from the S sub-ensemble estimates to the relevant observation, $\hat{y}_{j,i}$, by calculating their mean value:

$$\hat{y}_{e,i} = \frac{1}{S}\sum_{j=1}^{S}\left(\hat{y}_{j,i}\right). \tag{13}$$

The median statistic of the sub-ensemble estimates is defined as follows:

$$\hat{y}_{e,i} = median\left(\hat{y}_{1,i}, \hat{y}_{2,i}, \dots, \hat{y}_{k,i}\right). \tag{14}$$

The median is a robust tool that is not influenced by outlier values. Therefore, the median statistic reduces the influence of poor estimation performance related to some ensemble members. Note that sub-models that yield an under/over estimation in some cases may yield good estimates in other cases. Such occurrence is treated by utilizing the median statistic to achieve nonlinear ensemble integration.

The integration technique in this work applies a linear regression function on the training's sub-model estimates in order to generate a value for an ensemble estimate. Also, a linear regression function based on an ordinary least square (OLS) algorithm results in an efficient estimate of the regression parameters (Nelder and Wedderburn, 1972, Charnes et al., 1976, Stigler, 1986), where the expression is represented as:

$$\hat{y}_{e,i} = B_o + \sum_{j=1}^{k} \left(B_j \times \hat{y}_{j,i} \right), \tag{15}$$

where B_o and B_j's are the unknown linear regression coefficients that are computed by applying the OLS formulation on all the sub-models' estimates in the training stage. The coefficients of the multiple linear regression (MLR) are calculated analytically (Draper et al., 1966, Neter et al., 1996, Montgomery et al., 2012).

The Gaussian-distributed estimates will perform well due to employing the linear regression. The OLS estimates remove outlier estimates to a certain extent, but are affected by them. The advantage of using an OLS-based linear regression is that it allows for evaluating the performance of linear combiners in ensemble modeling. The parameters of the linear combiners are fixed and inferred, while taking into account all the estimates calculated in all sub-models. Hence, for each observation, all the related sub-models' estimates are combined into one ensemble estimate. Furthermore, a robust fit of the sub-models' estimates is performed (Andrews, 1974, Meer et al., 1991, Dumouchel and O'Brien, 1991, Holland and Welsch, 1977, Fox, 2002). This method results in robust estimates of the MLR coefficients. The proposed algorithm utilizes an iterative method based on least squares algorithm that is supported by a bi-square re-weighing function. It is known that the robust fitting technique requires a weighing

function. It also demands a tuning constant in order to arrive at a residual vector which is changed iteratively. The distribution of errors may be non-normal which is an inevitable problem. Such issue could be avoided by using robust regression or robust MLR techniques (Meer et al., 1991, Maronna et al., 2006, Huber et al., 1996). In this work, the robust regression method is used in the MATLAB environment. Such method generates the ensemble output, $\hat{y}_{e,i}$, based on the k sub-ensemble estimates, $\hat{y}_{j,i}$, as per the formula below:

$$\hat{y}_{e,i} = B_{robust_o} + \sum_{j=1}^{k} \left(B_{robust_j} \times \hat{y}_{j,i} \right), \tag{16}$$

where B_{robust_o} refers to the robust regression bias while B_{robust_j} refers to the robust regression coefficients. The coefficients are computed as the $(n+1)^{th}$ iteration which is the solution of a robust multi-linear regression. An iterative weighted least square function is computed as follows:

$$\sum_{i=1}^{N} \left(r_{i_{n+1}}\right)^2 = \sum_{i=i}^{N} w_{i_n} \left[\left[y_i - \left(B_{robust_{o_{n+1}}} + \sum_{j=1}^{k} \left(B_{robust_{j_{n+1}}} \times \hat{y}_{j,i} \right) \right) \right] \right]^2, \tag{17}$$

where $r_{i_{n+1}}$ is the $(n+1)^{th}$ weighted error function combining the individual model estimates on the i^{th} observation, $w_{i_{n+1}}$ refers to the assigned weight, y_i refers to the i^{th} observation obtained from training study, and $\hat{y}_{j,i}$ refers to the corresponding estimate, generated from the j^{th} individual model. Furthermore, the weights in the previous relation are updates as follows:

$$w_{i_{n+1}} = \left(\left| r_{i_{n+1}} \right| < 1 \right) \times \left(1 - \left(r_{i_{n+1}} \right)^2 \right)^2. \tag{18}$$

Also, the weighted residuals are updated based on the following process:

$$r_{i_{n+1}} = \left[\left(\frac{r_{i_n}}{tune \times \sigma_n}\right) \times \sqrt{(1 - e_i)}\right], \qquad \sigma_n = \frac{MAD_n}{0.6745}, \qquad (19)$$

where r_{i_n} refers to the model error on the corresponding i^{th} basin from the previous iteration, e_i is the leverage error value for the i^{th} observation using the OLS regression during the training stage. $tune$ is a scalar which significantly drives the degree of outlier influence on estimate of the robust coefficients, σ_n is the estimated deviation of the residuals, obtained from previous iteration. Moreover, MAD_n is the median absolute deviation of the residuals, obtained from previous iteration.

In this study, the tuning constant is fixed at 4.685. As a result, the coefficient estimates are 95% as statistically efficient as the ordinary least-squares estimates (Maronna et al., 2006). This parameter value is considered under the assumption that the response variable resembles a normal distribution, and that it has no outliers. Consequently, an increase in the tuning constant will magnify the influence of large residuals. Note that the value 0.6745 makes the estimate unbiased. The idea to include a robust fitting tool is inspired by the fact that certain sub-models in the ensemble generate outlier estimates continuously. Therefore, a result based on a median combiner only may lead to an incorrect choice of the ensemble estimate which is also based on the ensemble size, as well as the number of exaggerating models in the ensemble. Thus, the robust fitting technique provides an ensemble estimate in the form of a robust linear combination of the sub-ensemble estimates. Therefore, robust regression presents an advantage over stacking by introducing a parameter primarily responsible for bias correction, besides the weighted sum of the ensemble member outputs. It is anticipated that such integration technique can be successfully applied to many data cases. Note that the amount of information available for training will

directly influence the generalization ability the combiner. However, this issue is addressed before deciding which robust fitting technique to use.

4. Case Study

The proposed ensemble model is based on information gathered from the hydrometric station network which covers the southern part of the province of Québec, Canada. Winter and summer quantiles are examined separately due to the inconsistency in the low-flow generating phenomena. Since the dataset represent various sites in the province of Quebec, each site experience low-flow regimes due to notably different variations in the hydrologic process causing the low-flow. This process is naturally assumed unknown and the empirical model attempts to estimate its end-product of interest, i.e. the low flow. If the mechanism is the same, then the model will be able to exactly capture the low-flow quantiles in all the sites. The low-flow quantiles, with return periods of T of 2, 5 and 10 years and duration d of 7 and 30 days, are estimated in this work. These quantiles are of interest for fish habitat protection and water quality control (Ouarda and Shu, 2009). Moreover, in Canada, these quantiles are the most common indices for water supply system analysis during droughts as well as the studies of stream-based waste assimilation capacity (Ouarda et al., 2008a).

In this work, seven physiographical and meteorological variables are selected to study the seasonal low-flow quantiles. The variables are as follows, basin area (*A*), percent of basin covered by forest (*PFOR*), percent of basin being lake (*PLAKE*), annual mean degree days less than 0°c (*DJBZ*), annual mean days with temperature above 27°c (*NJH27*), summer mean liquid precipitation (*PLME*), and curve number (*CN*), which is a soil characteristic. Table 2 provides a summary of the descriptive statistics of all the study variables. *NJH27* is associated to a particular

regional hydrology and climatology benchmark for Québec (27°c), representing the medium temperature for the month of July based on maximum temperatures. Further details on the resources are available in (Ouarda et al., 2005).

Table 2: Descriptive statistics of explanatory variables.

Variable	Symbol	Mean	Max	Min	Standard Deviation
Basin area (Km^2)	*A*	5,655.52	96,600	0.70	11,685.70
Basin's area fraction occupied by lakes (%)	*PLAKE*	6.33	32.00	0.00	6.57
Basin's area fraction occupied by forest (%)	*PFOR*	85.78	100.00	6.50	15.97
Annual mean degree days < 0^oC (degree day)	*DJBZ*	1,635.15	2,963.10	920.60	529.29
Summer mean liquid precipitation (mm)	*PLME*	464.51	664.00	306.00	77.40
Average number of days with temperature > 27^oC	*NJH27*	12.28	36.60	0.80	7.57
Curve Number	*CN*	45.08	78.20	21.00	-

Initially, catchments corresponding to a network of 190 hydrometric stations are considered. To ensure the quality of the database, the stations should adhere the criteria below (Ouarda et al., 2005):

1- More than 10 years of flow record should be available.

2- The flow record of the station should be stationary (Kendall, 1975).

3- The flow record should be independent (Wald and Wolfowitz, 1943).

4- The assessed catchment should represent a natural flow incidence.

The above criteria resulted in a network of 129 sites, represented by their corresponding at-site stations, for the summer low-flow quantiles, and 135 and 133 sites for the winter low-flow quantiles with durations of 30-days and 7-days, respectively. This work considers catchments located between 45N and 55N longitude, and between 55W and 80W Latitude. In addition, the total area of each site ranges from 572 km^2 and up to 96,600 km^2. The seasonal low-flow quantiles are selected for return periods of 2, 5 and 10 years. Table 3 presents the correlations between the quantiles and the physiographic and meteorological variables. Further details about the summer and winter quantiles, the statistical approach to their at-site frequency analysis as well as the map of the sites are available in (Ouarda et al., 2005, Herrera-Guzman, 2008, Ouarda and Shu, 2009).

Table 3: Correlation between explanatory and response variables.

Variable	Summer Season			Winter Season		
	$Q_{5,30}$	$Q_{2,7}$	$Q_{10,7}$	$Q_{5,30}$	$Q_{2,7}$	$Q_{10,7}$
A	0.941	0.944	0.927	0.981	0.983	0.975
PLAKE	0.531	0.541	0.530	0.588	0.585	0.583
PFOR	-0.029	-0.031	-0.031	-0.074	-0.066	-0.067
DJBZ	0.575	0.572	0.566	0.558	0.585	0.583
NJH27	-0.344	-0.341	-0.343	-0.308	-0.301	-0.298
PLME	-0.432	-0.429	-0.426	-0.429	-0.428	-0.425
CN	-0.203	-0.214	-0.212	-0.173	-0.183	-0.181

5. Experimental Setup

A two-stage resampling-based ANN ensemble model is proposed to generalize the underlying relationship between physiographical and metrological variables, and hydrologic variables. Information homogeneity, or information mixture, and combiner-information parameters are disclosed in this work. The former is used to examine and smooth out information diversity discovered among ensemble members. The latter is used to evaluate performance and assess the sensitivity of linear combiners toward available information.

Initially, a pre-processing of inputs and outputs is carried out. The variables are first normalized; a linear scaling of each utilized variable is utilized so that the instances are bound between -1 and 1 to (Bishop, 2006). To optimize the ensemble configuration, different ensemble sizes, mixture ratios, dedicated data to combiners' training and ANN structures are investigated using a Jackknife validation approach (Ouarda and Shu, 2009). The two-stage resampling is applied on various ensemble size cases, after filtering the data for combiner training (see approach in Section 2.1). Different ANN configurations (as sub-ensembles) are investigated; in each jackknife validation study, an ensemble model's performance is assessed for each m_e, $m_{combiner}$ and combiner choice for the homogeneous ensemble where all the ensemble members are of the same structure.

The proposed validation approach aims to examine the relative performances of the regional low-flow estimation models (Charron and Ouarda, 2015). The quantile values are temporarily excluded from the database. The remaining sites are trained using the ensemble members and the ensemble combiners. Then, regional estimates can be collected for the ungauged site using the calibrated ensemble model. At-site estimates are later examined against ensemble quantile estimates for ungauged sites. The utilized predefined evaluation criteria are explained later in this section. This validation approach to the ANN ensemble members and to the ensemble model

itself arises from the fact that the proposed ensemble model's framework prompts specific conditions on the ANN members. Because the proposed ensemble framework uses explicit diversity parameters, validating the optimal ANN in an individual manner (as it is the case in conventional machine learning practice) does not make sense. This will prompt using all the training data. However, validating the ensemble model as a whole (which is slightly more complex but still computationally efficient given the parallel architecture and the simplified diversity evolution) should be considered. In other words, validating the overall performance of the ensemble model will be better than validating one specific ANN (using cross-validation approach). An ensemble validation approach helps examining the real performance of the ANN members, each, under the ensemble parameters' influence. Jackknife validation has a built-in sensitivity analysis that assesses the relationship between the model's performance and available information for training (Efron, 1981, Ouarda and Shu, 2009). Performance criteria include root mean square error (*RMSE*), relative root mean square error (*rRMSE*), bias (*Bias*), and relative bias (*rBias*). They measure the generalization ability for various ensemble sizes. Similarly, various homogeneity ratios are considered. Normalizing the error magnitude is an important step to accurately find *rBias* and *rRMSE*. The four measures are computed as per the equations below:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Q_{d,T_i} - \hat{Q}_{d,T_i}\right)^2}\ , \tag{20}$$

$$rRMSE = 100 \times \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{Q_{d,T_i} - \hat{Q}_{d,T_i}}{Q_{d,T_i}}\right)^2}\ , \tag{21}$$

$$Bias = \frac{1}{n} \times \sum_{i=1}^{n} \left(Q_{d,T_i} - \hat{Q}_{d,T_i} \right) , \quad (22)$$

$$rBias = \frac{100}{n} \times \sum_{i=1}^{n} \left(\frac{Q_{d,T_i} - \hat{Q}_{d,T_i}}{Q_{d,T_i}} \right) , \quad (23)$$

where Q_{d,T_i} is at-site *d*-day, *T*-year drought quantile value of site i, $\hat{Q}_{d,T_i}$ is the corresponding estimate from the final ensemble learner, and *n* is the sample size of the validation set of observations (or sites).

The performance of the proposed model is examined based on six low-flow quantiles, the summer season corresponds for three quantiles while the winter season corresponds for the other three quintiles being $Q_{7,2}$, $Q_{7,5}$ and $Q_{30,5}$. The results of the work by (Ouarda and Shu, 2009) are used as a benchmark for evaluating the proposed method. Jackknife trials are simulated for each low-flow quantile. These simulations evaluate various combinations of homogeneity parameters, ANN members' structure, ensemble sizes, and ensemble combiners, to determine the optimum ensemble model. The validation results are discussed in the next section. Also, the performance of the optimal ensemble models is assessed against the benchmark study. It is important to note that the Jackknife validation is the testing segment of the study. In Jackknife validation, one instance from the available dataset is blocked, and the remaining instances are used for training. The omitted instance is then used for testing. This is repeated until all instances in the available dataset are tested, and the testing performance is reported. The training performance is not reported in the manuscript as it is not an indication of the generalization ability, and the sub-models are already regularized via early stopping and internal six fold cross-validation during their training phase.

6. Results and Discussion

Optimum ensemble configuration is selected for each low-flow quantile in the winter and summer seasons. Jackknife validation results aim at assessing the models' generalization ability over the available observations. The proposed models' performance is compared to that of the benchmark model.

Figures 3 and 4 preview the jackknife validation performance, with respect to mixture level in sub-sample, of selected ensemble models for the low-flow quantiles in the winter and summer seasons, respectively. For each season-based quantile, the selected ensemble models incorporate ANN members with the same complexity. The variables in the figures are the combiner choice and the homogeneity levels. The four ensemble models show a similar trend in performance sensitivity with respect to different homogeneity levels. However, in all six quantiles, the ensemble models with mean and median combiners are relatively more stable along different homogeneity levels.

The OLS and robust fitting tools adopt a relatively more sensitive performance at low homogeneity levels. This behavior can be attributed to the nature of the sub-ensemble estimation behavior, which will be more variable at such mixture levels. Furthermore, the original sample, in this case study, has a small number of observations to be used in the ensemble training phase. At high homogeneity levels, the sensitivity in jackknife performance of ensembles with OLS and robust fitting tools is similar to that of ensembles with mean and median combiners. This behavior is due to the fact that more training information is shared between the sub-ensembles and, hence, the estimation behavior tends to be comparable for all the members. The number of hidden neurons in the ANN members represents the complexity of the link between the inputs

and outputs in the proposed system. The mixture levels can be regarded as a more precise measure of model generalization ability, other than the Jackknife validation error. To this extent, the model's generalization ability is now further analyzed through its diversity behavior (amount of information required by the ensemble members to produce the optimum behavior). In Table 4, the optimum configuration for the ensemble models to estimate the low-flow quantiles in the winter and summer seasons is presented. For the summer season, it is shown that the optimum ANN structures for $Q_{2,7}$ and $Q_{10,7}$ ensemble models require nine hidden neurons, while $Q_{5,30}$ ensemble model requires eleven hidden neurons.

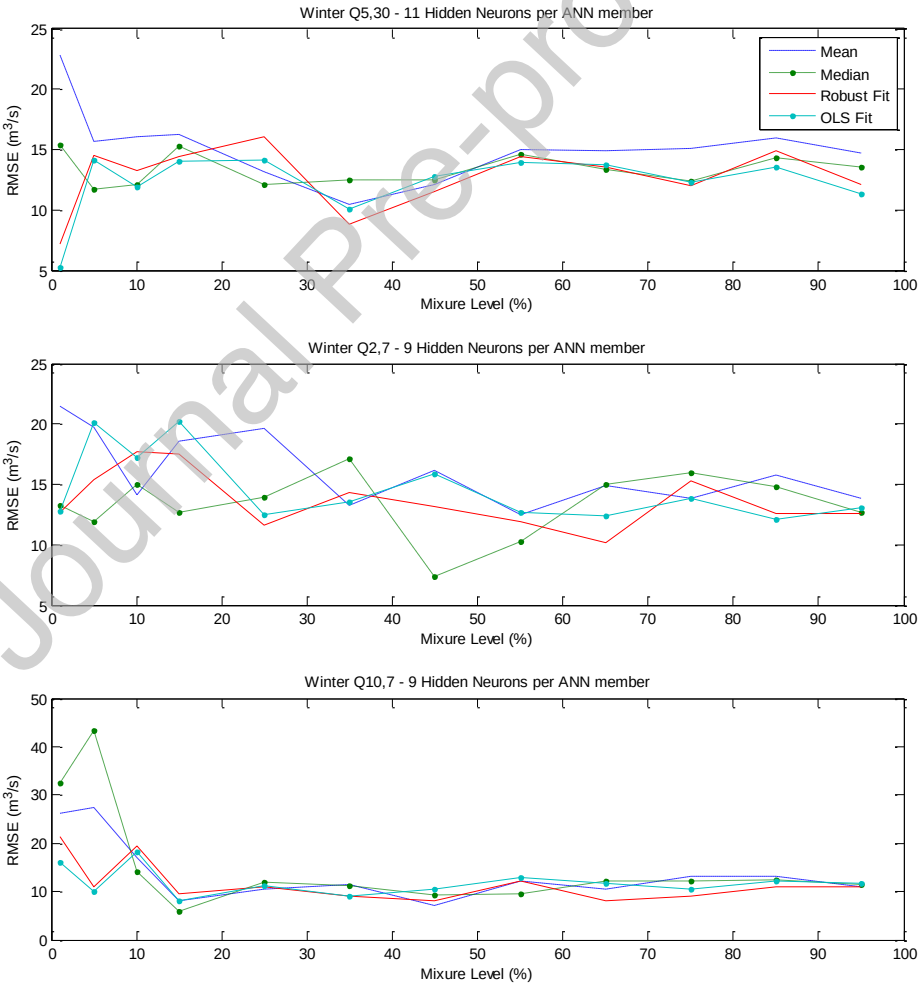

Figure 3: Jackknife results for selected ensembles to the winter study with respect to homogeneity levels.

For the winter season, ANNs require twelve hidden neurons for ensemble models estimating $Q_{2,7}$ and $Q_{10,7}$, and eleven hidden neurons for $Q_{5,30}$ ensemble model. It is also shown that, for both seasons, the $Q_{5,30}$ ensemble models incorporated ANN members with similar complexity (number of hidden neurons) and m_c value, but they adopted different m_e values and different ensemble combiner choices. The ANN members for the ensembles explaining $Q_{2,7}$ and $Q_{10,7}$ quantiles are shown to have the same complexity in the same season, with the summer models being more complex. It should be noted that in the benchmark study, the optimal ANN configuration for summer and winter quantiles are different than in the present study. This is expected due to the employed ensemble validation approach, as discussed earlier.

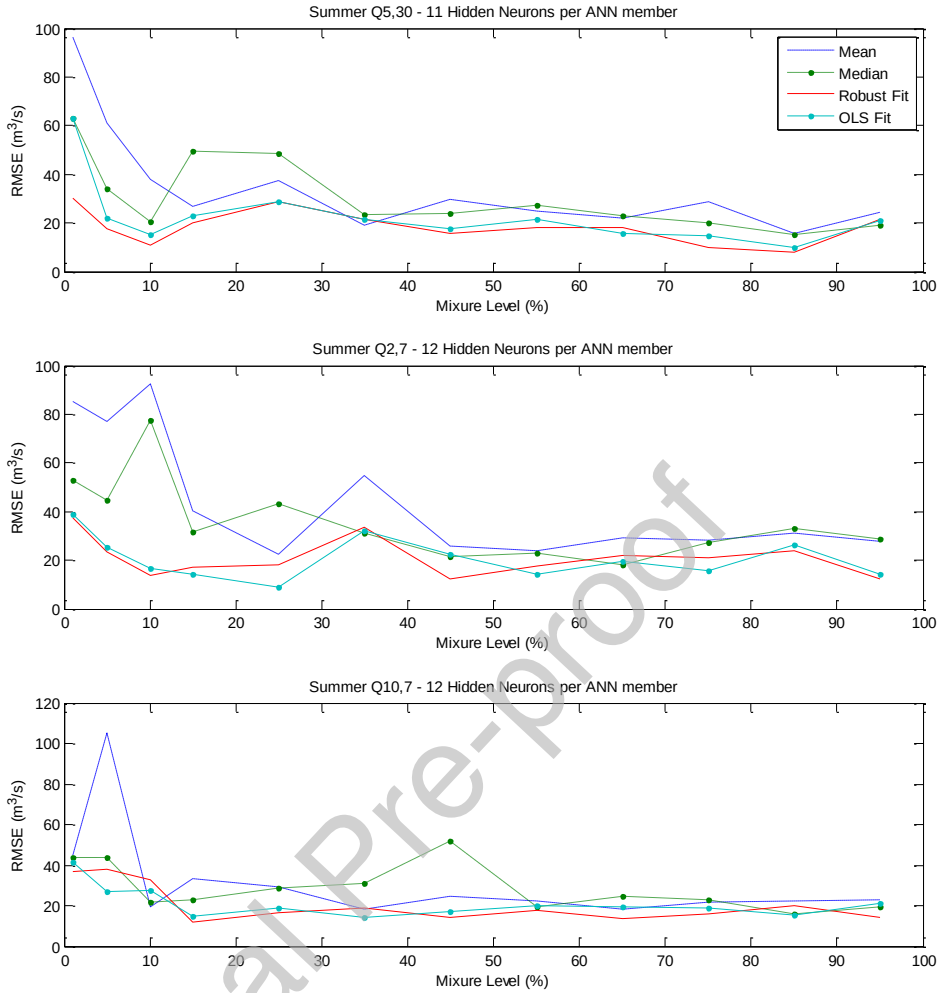

Figure 4: Jackknife results for selected ensembles to the summer study with respect to homogeneity levels.

The similar complexity in $Q_{2,7}$ and $Q_{10,7}$ ensembles validates the adequacy of the optimum models. The low-flow duration in a given season is of the same number of days for both $Q_{2,7}$ and $Q_{10,7}$. Consequentially, the regression models for these quantiles are expected to have the same complexity. Furthermore, all optimum ensemble models incorporated m_e values less than 1, meaning that the ANN models did not require all available information for learning to produce best models, and that diverse-based ensemble models produced better generalization ability. This

finding confirms the benefit of the proposed ensemble framework as a diversity promoting system. In addition, validation of the ensemble studied models with different homogeneity levels emphasize on the importance of the diversity control among the ensemble members.

Table 4: Optimum ensemble configuration for low-flow quantiles in the winter and summer seasons.

Optimum Ensembles for Low-Flow Quantiles In The Winter Season					
Quantile	Ensemble Size	Hidden Neurons	m_e (%)	m_c (%)	Combiner
$Q_{5,30}$	5	11	1	5	OLS Fit
$Q_{2,7}$	5	9	10	5	OLS Fit
$Q_{10,7}$	5	9	10	5	Robust Fit
Optimum Ensembles for Low-Flow Quantiles In The Summer Season					
Quantile	Ensemble Size	Hidden Neurons	m_e (%)	m_c (%)	Combiner
$Q_{5,30}$	5	11	85	5	Robust Fit
$Q_{2,7}$	5	12	25	5	OLS Fit
$Q_{10,7}$	10	12	5	10	OLS Fit

All optimum ensemble models incorporated m_c values above 0. The optimized performance of the selected regression-based combiners requires observations that are not used in the sub-ensembles' training. This is due to the over-fitting consequence of observations used in the training of the sub-ensembles. If all information is used in training, the linear combiners will not add to the accuracy of the ensemble estimate. However, when a portion of observations is kept away from the ensemble members, the estimates of such observations will not be over-fit and will guide the training of the combiners more properly. It is important to note the dynamic tradeoff between the distribution of the information and the performance of the ensemble models. If more observations are kept out of the resamples (higher m_c values), the improvement in the combiner's

generalization ability will be met by deterioration in sub-ensembles' experience, and vice-versa. The optimized homogeneity levels are, as a consequence, dependent on the amount of available information and should be selected with consideration to the discussed tradeoff.

Table 5 presents a comparison of the Jackknife validation results between the benchmark models and the proposed optimum models and configuration for estimating the low-flow quantiles in the winter and summer seasons. The proposed ensemble model has significantly improved over the benchmark results; for example, for the winter-season $Q_{5,30}$ model, the *RMSE* dropped from 15.84 (m^3/s) to 5.24 (m^3/s), the *Bias* error dropped from 1.61 (m^3/s) to 0.25 (m^3/s), *rRMSE* dropped from 34.87% to 26.07%, and *rBIAS* dropped from -5.13% to -0.81%. The improvement in the absolute and relative error measures indicates that not only the higher values in the quantile space became better estimated, rather than the case of underestimation by the benchmark models, but also the estimates of lower values improved. Hence, the adverse scale problem in regional extreme event estimation is further treated by the proposed models. Because the ANN members will always have certain estimation accuracy around the real value, this can be attributed to the combiners chosen for the proposed ensemble model.

A distinct example is shown in the winter $Q_{2,7}$ and $Q_{10,7}$ optimum ensembles, where the OLS linear combiner and robust fitting combiner are selected, respectively. The choice of robust fitting as a combiner in the optimum winter $Q_{10,7}$ model is expected to be a result of the nature of the target variable itself, as it reside in a more extreme location at the tail of the distribution than the winter $Q_{2,7}$ variable (Ouarda and Shu, 2009). The linear regression techniques used as combiners enjoy the bias correcting parameters that, in contrast to stacking approaches, directly target bias reduction. This is notably present in the results of the proposed models for all the low-flow quantiles, where bias error is significantly reduced. It is also shown that the optimum m_c value is

above 0% for all models; this means that the optimum models required unique information for the training of the combiners' parameters even though the ANN-combiner training data tradeoff occurred. Figures 4 and 6 show the improvement of the estimation accuracy of all the low-flow quantiles in the winter and the summer seasons, respectively. From the figures, the inferiority of scale challenge and the discrepancy in some of the at-site low-flow observations have been mitigated.

Table 5: Comparison of Jackknife validation results between the benchmark and the proposed models for the estimation of winter and summer low-flow quantiles.

Winter Season					
Quantile	Reference	*RMSE* (m^3/s)	*rRMSE* (%)	*Bias* (m^3/s)	*rBias* (%)
$Q_{5,30}$	Benchmark	15.84	34.87	1.61	-5.13
	Proposed Approach	**5.24**	**26.07**	**0.25**	**-0.81**
$Q_{2,7}$	Benchmark	16.59	33.13	1.66	-4.55
	Proposed Approach	**7.72**	**16.80**	**0.49**	**-2.23**
$Q_{10,7}$	Benchmark	13.91	42.92	1.10	-6.87
	Proposed Approach	**5.29**	**22.91**	**0.06**	**-1.98**
Summer Season					
Quantile	Reference	*RMSE* (m^3/s)	*rRMSE* (%)	*Bias* (m^3/s)	*rBias* (%)
$Q_{5,30}$	Benchmark	27.95	31.02	0.94	-3.08
	Proposed Approach	**7.99**	**25.77**	**0.37**	**-1.74**
$Q_{2,7}$	Benchmark	35.90	31.41	5.47	-1.65
	Proposed Approach	**8.78**	**23.39**	**0.45**	**-1.59**
$Q_{10,7}$	Benchmark	27.33	39.17	2.69	-3.17
	Proposed Approach	**7.41**	**33.49**	**0.18**	**-2.71**

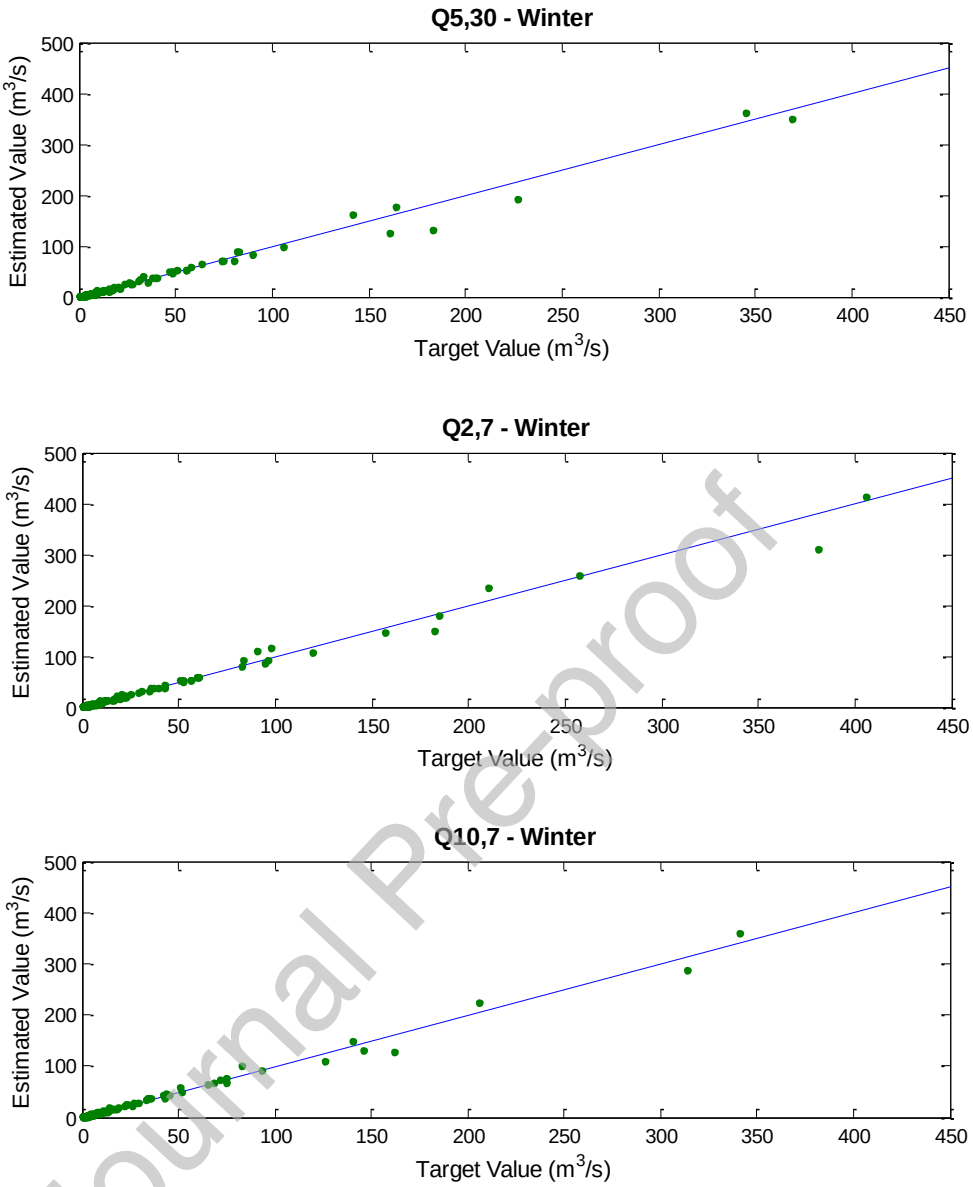

Figure 5: Scatter plot of Jackknife validation results of winter low-flow quantiles.

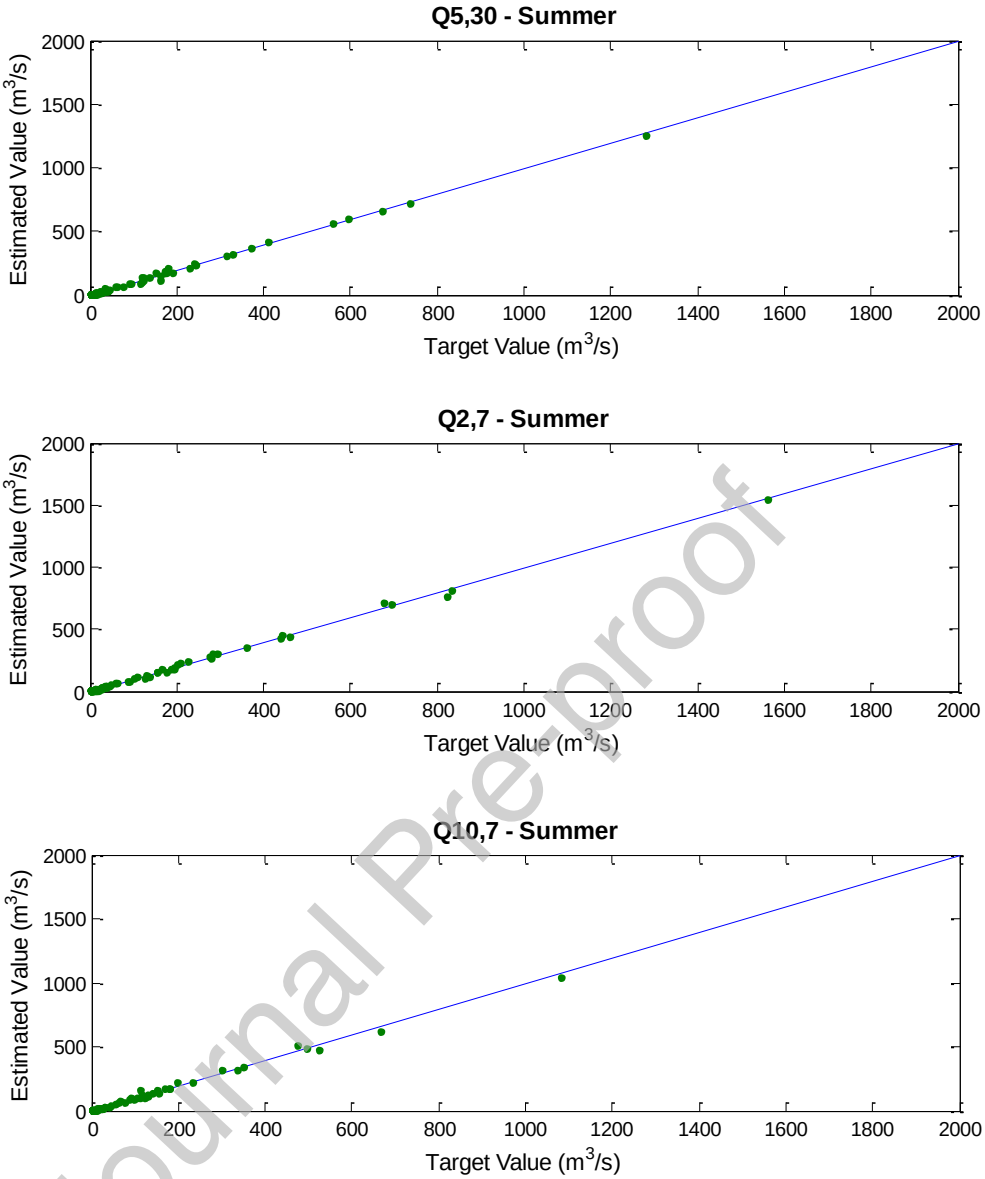

Figure 6: Scatter plot of Jackknife validation results of summer low-flow quantiles.

7. Conclusions

An ensemble framework is proposed in this study to improve the generalization ability of the regression sub-ensemble models. The proposed two-stage resampling algorithm makes use of the homogeneity concept between subsamples. Further, the idea of isolating some of the training data from the sub-model training and using it in the combiner training has enhanced the performance of the corresponding optimum ensemble models. The over-fitting disadvantage of neural networks in ensemble modeling has been treated using the proposed resampling plan. The results clearly show substantial enhancement in the estimation accuracy. The magnitudes of the homogeneity levels corresponding to the optimum ensemble models have indeed promoted the diversity concept in the theory of ensemble learning theory; the optimum mixture levels are found to be low enough, although significant, for ensemble members to be trained using sub-sample with diverse information about the system. It is shown that the sensitivity of the combiners' performance is indirectly related to the mixture levels, through the level of diversity in the individual members.

The generalization ability of any predictor in the field of regional frequency analysis relies on the relationship between the hydrologic stations considered in that study. Hence, understanding the homogeneity, or hydrologic similarity, between the stations is expected to significantly improve the regression models over them. In fact, assessing and modeling the level of homogeneity within a group of stations is an active topic of research (Chebana and Ouarda, 2007). Future work may consider modeling the level of hydrologic homogeneity in the data first and then investigate the data-mixing scheme as a function of the hydrologic homogeneity. Such approach opens the door to integrate the concept of diversity within the physical identity of the system of interest, providing solutions to limitations in the current ensemble models for the

challenging low-flow quantile estimation task. Hence, future work may also target the application of ensembles involving different mixture ratio relations, given a relatively large database, and develop a validation approach to such models. A nonlinear mixture relation of the information share between the ensemble members could further optimize the generalization ability of such ensembles. The possible combinations between the members can easily reach a relatively huge number. It is costly to investigate all possible combinations for means of model validation and selection. To alleviate the difficulty in such work, search-based optimization algorithms may be investigated to minimize the number of simulations mandatory for finding the optimum ensemble structure.

Author Statement

Mohammad H Alobaidi: Conceptualization, Methodology, Data curation, Software, Formal analysis, Writing- Original draft preparation. **Taha B.M.J. Ouarda:** Visualization, Resources, Writing- Reviewing and Editing, Supervision. **Prashanth R. Marpu:** Writing- Reviewing and Editing, Data curation, Investigation, Validation. **Fateh Chebana:** Writing-Reviewing and Editing, Validation.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

AGRAFIOTIS, D. K., CEDENO, W. & LOBANOV, V. S. 2002. On the use of neural network ensembles in QSAR and QSPR. *Journal of chemical information and computer sciences,* 42**,** 903-911.

AJAMI, N. K., DUAN, Q., GAO, X. & SOROOSHIAN, S. 2006. Multimodel Combination Techniques for Analysis of Hydrological Simulations: Application to Distributed Model Intercomparison Project Results. *Journal of Hydrometeorology,* 7**,** 755-768.

ALAM, K. M. R., SIDDIQUE, N. & ADELI, H. 2019. A dynamic ensemble learning algorithm for neural networks. *Neural Computing and Applications***,** 1-16.

ALOBAIDI, M. H., MARPU, P. R., OUARDA, T. B. & CHEBANA, F. 2015. Regional frequency analysis at ungauged sites using a two-stage resampling generalized ensemble framework. *Advances in water resources,* 84**,** 103-111.

ALOBAIDI, M. H., MARPU, P. R., OUARDA, T. B. & GHEDIRA, H. 2014. Mapping of the solar irradiance in the UAE using advanced artificial neural network ensemble. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing,* 7**,** 3668-3680.

ALOBAIDI, M. H., MEGUID, M. A. & CHEBANA, F. 2019. Predicting seismic-induced liquefaction through ensemble learning frameworks. *Scientific reports,* 9**,** 1-12.

ANDREWS, D. F. 1974. A robust method for multiple linear regression. *Technometrics,* 16**,** 523-531.

BASU, B. & SRINIVAS, V. 2014. Regional flood frequency analysis using kernel-based fuzzy clustering approach. *Water Resources Research,* 50**,** 3295-3316.

BISHOP, C. M. 2006. *Pattern recognition and machine learning,* springer New York.

BREIMAN, L. 1996a. Bagging predictors. *Machine learning,* 24**,** 123-140.

BREIMAN, L. 1996b. Stacked regressions. *Machine learning,* 24**,** 49-64.

BREIMAN, L. 2001. Random forests. *Machine learning,* 45**,** 5-32.

BROWN, G. 2004. *Diversity in neural network ensembles.* U191868 Ph.D., University of Birmingham (United Kingdom).

BROWN, G., WYATT, J., HARRIS, R. & YAO, X. 2005a. Diversity creation methods: a survey and categorisation. *Information Fusion,* 6**,** 5-20.

BROWN, G., WYATT, J. L. & TI, P. 2005b. Managing Diversity in Regression Ensembles. *J. Mach. Learn. Res.,* 6**,** 1621-1650.

BÜHLMANN, P. & HOTHORN, T. 2007. Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science***,** 477-505.

BÜHLMANN, P. L. Bagging, subagging and bragging for improving some prediction algorithms. Seminar für Statistik, 2003 Eidgenössische Technische Hochschule (ETH), Zürich, Switzerland. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH), Zürich.

CHARNES, A., FROME, E. & YU, P.-L. 1976. The equivalence of generalized least squares and maximum likelihood estimates in the exponential family. *Journal of the American Statistical Association,* 71**,** 169-171.

CHARRON, C. & OUARDA, T. B. M. J. 2015. Regional low-flow frequency analysis with a recession parameter from a non-linear reservoir model. *Journal of Hydrology,* 524**,** 468-475.

CHEBANA, F. & OUARDA, T. B. M. J. 2007. Multivariate L-moment homogeneity test. *Water Resources Research,* 43**,** W08406.

CHEN, H., COHN, A. G. & YAO, X. 2012. Ensemble Learning by Negative Correlation Learning. *Ensemble Machine Learning: Methods and Applications***,** 177.

CLARKE, B. 2003. Comparing Bayes model averaging and stacking when model approximation error cannot be ignored. *Journal of Machine Learning Research,* 4**,** 683-712.

DEMUTH, H. B., BEALE, M. & HAGAN, M. 2006. *Neural Network Toolbox for Use with MATLAB: User's Guide*, MathWorks, Incorporated.

DIETTERICH, T. 2000. Ensemble methods in machine learning. *Multiple classifier systems***,** 1-15.

DIKS, C. H. & VRUGT, J. 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. *Stochastic Environmental Research and Risk Assessment,* 24**,** 809-820.

DINGMAN, S. L. & LAWLOR, S. C. 1995. ESTIMATING LOW-FLOW QUANTILES FROM DRAINAGE-BASIN CHARACTERISTICS IN NEW HAMPSHIRE AND VERMONT1. *JAWRA Journal of the American Water Resources Association,* 31**,** 243-256.

DONG, L., XIONG, L. & YU, K.-X. 2013. Uncertainty analysis of multiple hydrologic models using the Bayesian model averaging method. *Journal of Applied Mathematics,* 2013.

DONG, X., YU, Z., CAO, W., SHI, Y. & MA, Q. 2020. A survey on ensemble learning. *Frontiers of Computer Science***,** 1-18.

DRAPER, N. R., SMITH, H. & POWNELL, E. 1966. *Applied regression analysis*, Wiley New York.

DRUCKER, H. Improving regressors using boosting techniques. MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-, 1997. Citeseer, 107-115.

DUAN, Q., AJAMI, N. K., GAO, X. & SOROOSHIAN, S. 2007. Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources,* 30**,** 1371-1386.

DUFFY, N. & HELMBOLD, D. 2002. Boosting methods for regression. *Machine Learning,* 47**,** 153-200.

DUMOUCHEL, W. & O'BRIEN, F. 1991. Integrating a robust option into a multiple regression computing environment. *Computing and graphics in statistics.* Springer-Verlag New York, Inc.

EFRON, B. 1981. Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika,* 68**,** 589-599.

EFRON, B. 1982. *The jackknife, the bootstrap, and other resampling plans*, Siam.

ERDAL, H. I. & KARAKURT, O. 2013. Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms. *Journal of Hydrology,* 477**,** 119-128.

FOX, J. 2002. Robust Regression. *Appendix to An R and S-PLUS Companion to Applied Regression.*

FRANCKE, T., LÓPEZ-TARAZÓN, J. A. & SCHRÖDER, B. 2008. Estimation of suspended sediment concentration and yield using linear models, random forests and quantile regression forests. *Hydrological Processes,* 22**,** 4892-4904.

FREUND, Y. & SCHAPIRE, R. E. Experiments with a new boosting algorithm. MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-, 1996. MORGAN KAUFMANN PUBLISHERS, INC., 148-156.

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. 2000. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *The annals of statistics,* 28**,** 337-407.

FRIEDMAN, J. H. 2001. Greedy function approximation: a gradient boosting machine.(English summary). *Ann. Statist,* 29**,** 1189-1232.

GEMAN, S., BIENENSTOCK, E. & DOURSAT, R. 1992. Neural networks and the bias/variance dilemma. *Neural computation,* 4**,** 1-58.

GOVINDARAJU, R. S. & RAO, A. R. 2010. *Artificial neural networks in hydrology*, Springer Publishing Company, Incorporated.

GRANITTO, P. M., VERDES, P. F. & CECCATTO, H. A. 2005. Neural network ensembles: evaluation of aggregation algorithms. *Artificial Intelligence,* 163**,** 139-162.

GREEN, M. & OHLSSON, M. Comparison of standard resampling methods for performance estimation of artificial neural network ensembles. The Third International Conference on Computational Intelligence in Medicine and Healthcare, July, 2007. 25-27.

GUSTARD, A. & DEMUTH, S. 2009. Manual on low-flow estimation and prediction. Opera.

HAGAN, M. T. & MENHAJ, M. B. 1994. Training feedforward networks with the Marquardt algorithm. *Neural Networks, IEEE Transactions on,* 5**,** 989-993.

HAGIWARA, K. 2002. Regularization learning, early stopping and biased estimator. *Neurocomputing,* 48**,** 937-955.

HANSEN, L. K. & SALAMON, P. 1990. Neural Network Ensembles. *IEEE Trans. Pattern Anal. Mach. Intell.,* 12**,** 993-1001.

HASHEM, S. 1993. *Optimal linear combinations of neural networks.* Ph.D., School of Industrial Engineering, University of Purdue.

HASHEM, S. 1997. Optimal Linear Combinations of Neural Networks. *Neural networks,* 10**,** 599-614.

HASHEM, S., SCHMEISER, B. & YIH, Y. Optimal linear combinations of neural networks: an overview. Neural Networks, IEEE World Congress on Computational Intelligence., 1994. IEEE, 1507-1512.

HERRERA-GUZMAN, E. 2008. *Développement d'une méthodologie hydrologique/statistique pour estimer les débits d'étiage au Québec habité.* Université du Québec, Institut national de la recherche scientifique.

HO, T. K. Random decision forests. Proceedings of 3rd international conference on document analysis and recognition, 1995. IEEE, 278-282.

HO, T. K. 1998. The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence,* 20**,** 832-844.

HOLLAND, P. W. & WELSCH, R. E. 1977. Robust regression using iteratively reweighted least-squares. *Communications in Statistics - Theory and Methods,* 6**,** 813-827.

HUBER, P. J., HUBER, P., STATISTICIEN, M., SUISSE, E. U. & STATISTICIAN, M. 1996. *Robust statistical procedures,* SIAM.

HUO, W., LI, Z., WANG, J., YAO, C., ZHANG, K. & HUANG, Y. 2019. Multiple hydrological models comparison and an improved Bayesian model averaging approach for ensemble prediction over semi-humid regions. *Stochastic Environmental Research and Risk Assessment,* 33**,** 217-238.

ISLAM, M. M., YAO, X. & MURASE, K. 2003. A constructive algorithm for training cooperative neural network ensembles. *Neural Networks, IEEE Transactions on,* 14**,** 820-834.

KENDALL, M. 1975. *Multivariate analysis*, Charles Griffin.

KOUIDER, A. 2003. *Analyse Fréquentielle Locale des Crues au Québec.* M.Sc., Institut National de la Recherche Scientifique.

KROGH, A. & VEDELSBY, J. 1995. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems***,** 231-238.

KUNCHEVA, L. I. That elusive diversity in classifier ensembles. Iberian conference on pattern recognition and image analysis, 2003. Springer, 1126-1138.

KUNCHEVA, L. I. & WHITAKER, C. J. 2003. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine learning,* 51**,** 181-207.

LÁZARO, M., HERRERA, F. & FIGUEIRAS-VIDAL, A. R. 2020. Ensembles of cost-diverse Bayesian neural learners for imbalanced binary classification. *Information Sciences,* 520**,** 31-45.

LEARNER, E. 1978. Specification Searches: Ad Hoc Inference l. *Wth Non*.

LIU, Y. 1999. *Negative correlation learning and evolutionary design of neural network ensembles.* 0800958 Ph.D., University of New South Wales (Australia).

MACLIN, R. & OPITZ, D. 1999. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research,* 11**,** 169-198.

MARONNA, R. A., MARTIN, R. D. & YOHAI, V. J. 2006. *Robust statistics*, J. Wiley.

MCCUEN, R., LEAHY, R. & JOHNSON, P. 1990. Problems with Logarithmic Transformations in Regression. *Journal of Hydraulic Engineering,* 116**,** 414-428.

MEER, P., MINTZ, D., ROSENFELD, A. & KIM, D. 1991. Robust regression methods for computer vision: A review. *International Journal of Computer Vision,* 6**,** 59-70.

MENDES-MOREIRA, J., SOARES, C., JORGE, A. M. & SOUSA, J. F. D. 2012. Ensemble approaches for regression: A survey. *ACM Computing Surveys (CSUR),* 45**,** 10.

MONTGOMERY, D. C., PECK, E. A. & VINING, G. G. 2012. *Introduction to linear regression analysis*, Wiley.

NELDER, J. A. & WEDDERBURN, R. W. 1972. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)***,** 370-384.

NETER, J., WASSERMAN, W. & KUTNER, M. H. 1996. *Applied linear statistical models*, Irwin Chicago.

OUARDA, T. B., CHARRON, C. & ST-HILAIRE, A. 2008a. Statistical models and the estimation of low flows. *Canadian Water Resources Journal,* 33**,** 195-206.

OUARDA, T. B. M. J., BÂ, K., DIAZ-DELGADO, C., CÂRSTEANU, A., CHOKMANI, K., GINGRAS, H., QUENTIN, E., TRUJILLO, E. & BOBÉE, B. 2008b. Intercomparison of regional flood frequency estimation methods at ungauged sites for a Mexican case study. *Journal of Hydrology,* 348**,** 40-58.

OUARDA, T. B. M. J., GIRARD, C., CAVADIAS, G. S. & BOBÉE, B. 2001. Regional flood frequency estimation with canonical correlation analysis. *Journal of Hydrology,* 254**,** 157-173.

OUARDA, T. B. M. J., JOURDAIN, V., GIGNAC, N., GINGRAS, H., HERRERA, H. & BOBÉE, B. 2005. Development of a hydrological model for the regional estimation of low-flows in the province of Quebec (in French). Eau, Terre, et Environ., Institut national de la recherche scientifique. *Res. Rep*.

OUARDA, T. B. M. J. & SHU, C. 2009. Regional low-flow frequency analysis using single and ensemble artificial neural networks. *Water Resources Research,* 45**,** W11428.

QU, B., ZHANG, X., PAPPENBERGER, F., ZHANG, T. & FANG, Y. 2017. Multi-model grand ensemble hydrologic forecasting in the Fu river basin using Bayesian model averaging. *Water,* 9**,** 74.

SCHMIDT, K. 2004. *Diversity in neural network ensembles.* 3117015 D.CS., Colorado Technical University.

SHARKEY, A. J. 1999. Boosting using neural networks. *Combining artificial neural Nets.* Springer.

SHU, C. & BURN, D. 2004. Artificial neural network ensembles and their application in pooled flood frequency analysis. *Water Resources Research,* 40**,** W09301.

SHU, C. & OUARDA, T. B. M. J. 2007. Flood frequency analysis at ungauged sites using artificial neural networks in canonical correlation analysis physiographic space. *Water Resources Research,* 43**,** W07438.

SHU, C. & OUARDA, T. B. M. J. 2008. Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system. *Journal of Hydrology,* 349**,** 31-43.

SIOU, L. K. A., JOHANNET, A., BORRELL, V. & PISTRE, S. 2011. Complexity selection of a neural network model for karst flood forecasting: The case of the Lez Basin (southern France). *Journal of Hydrology,* 403**,** 367-380.

SLAVIN ROSS, A., PAN, W., CELI, L. A. & DOSHI-VELEZ, F. 2019. Ensembles of Locally Independent Prediction Models. *arXiv***,** arXiv: 1911.01291.

SMAKHTIN, V. U. 2001a. Low flow hydrology: a review. *Journal of hydrology,* 240**,** 147-186.

SMAKHTIN, V. U. 2001b. Low flow hydrology: a review. *Journal of Hydrology,* 240**,** 147-186.

STIGLER, S. M. 1986. *The history of statistics: The measurement of uncertainty before 1900,* Harvard University Press.

SUN, T. & ZHOU, Z.-H. 2018. Structural diversity for decision tree ensemble learning. *Frontiers of Computer Science,* 12**,** 560-570.

TETKO, I. V. & VILLA, A. E. 1997. An enhancement of generalization ability in cascade correlation algorithm by avoidance of overfitting/overtraining problem. *Neural Processing Letters,* 6, 43-50.

THOMAS, D. M. & BENSON, M. A. 1970. *Generalization of streamflow characteristics from drainage-basin characteristics,* US Government Printing Office Washington, DC, USA.

TIKHONOV, A. & ARSENIN, V. Y. 1979. Methods of solution of ill-posed problems. Nauka, Moscow.

UEDA, N. & NAKANO, R. Generalization error of ensemble estimators. Neural Networks, 1996., IEEE International Conference on, 3-6 Jun 1996 1996. 90-95 vol.1.

VAPNIK, V. N. 1998. Statistical learning theory.

VOGEL, R. M. & KROLL, C. N. 1990. GENERALIZED LOW-FLOW FREQUENCY RELATIONSHIPS FOR UNGAGED SITES IN MASSACHUSETTS1. *JAWRA Journal of the American Water Resources Association,* 26**,** 241-253.

VOGEL, R. M. & KROLL, C. N. 1992. Regional geohydrologic-geomorphic relationships for the estimation of low-flow statistics. *Water Resources Research,* 28**,** 2451-2458.

VRUGT, J. A. & ROBINSON, B. A. 2007. Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resources Research,* 43**,** W01411.

WALD, A. & WOLFOWITZ, J. 1943. An exact test for randomness in the non-parametric case based on serial correlation. *The Annals of Mathematical Statistics,* 14**,** 378-388.

WOLPERT, D. H. 1992. Stacked generalization. *Neural networks*, 5**,** 241-259.
ZAIER, I., SHU, C., OUARDA, T. B. M. J., SEIDOU, O. & CHEBANA, F. 2010. Estimation of ice thickness on lakes using artificial neural network ensembles. *Journal of Hydrology*, 383**,** 330-340.
ZHANG, C. & MA, Y. 2012. *Ensemble Machine Learning*, Springer.
ZHOU, Z.-H. & CHEN, S. 2002. Neural network ensemble. *CHINESE JOURNAL OF COMPUTERS-CHINESE EDITION-*, 25**,** 1-8.