scientific reports

OPEN

Uncertainty quantification of granular computing-neural network model for prediction of pollutant longitudinal dispersion coefficient in aquatic streams

Behzad Ghiasi[1,10], Roohollah Noori[1,2,10]✉, Hossein Sheikhian[3,10], Amin Zeynolabedin[4], Yuanbin Sun[5], Changhyun Jun[6], Mohamed Hamouda[7]✉, Sayed M. Bateni[8] & Soroush Abolfathi[9]

Discharge of pollution loads into natural water systems remains a global challenge that threatens water and food supply, as well as endangering ecosystem services. Natural rehabilitation of contaminated streams is mainly influenced by the longitudinal dispersion coefficient, or the rate of longitudinal dispersion (D_x), a key parameter with large spatiotemporal fluctuations that characterizes pollution transport. The large uncertainty in estimation of D_x in streams limits the water quality assessment in natural streams and design of water quality enhancement strategies. This study develops an artificial intelligence-based predictive model, coupling granular computing and neural network models (GrC-ANN) to provide robust estimation of D_x and its uncertainty for a range of flow-geometric conditions with high spatiotemporal variability. Uncertainty analysis of D_x estimated from the proposed GrC-ANN model was performed by alteration of the training data used to tune the model. Modified bootstrap method was employed to generate different training patterns through resampling from a global database of tracer experiments in streams with 503 datapoints. Comparison between the D_x values estimated by GrC-ANN to those determined from tracer measurements shows the appropriateness and robustness of the proposed method in determining the rate of longitudinal dispersion. The GrC-ANN model with the narrowest bandwidth of estimated uncertainty (bandwidth-*factor* = 0.56) that brackets the highest percentage of true D_x data (i.e., 100%) is the best model to compute D_x in streams. Considering the significant inherent uncertainty reported in the previous D_x models, the GrC-ANN model developed in this study is shown to have a robust performance for evaluating pollutant mixing (D_x) in turbulent environmental flow systems.

Discharge of pollution loads into streams threatens the water and food supply, along with aquatic biodiversity at a global scale[1,2]. Natural rehabilitation of polluted streams is mainly characterized by the longitudinal dispersion coefficient, or the rate of longitudinal dispersion (D_x or K_x), a key parameter in river water quality models with large temporal and spatial variations. A challenging task in the study of the pollutant fate and transport in turbulent flow systems (e.g., streams) is determining D_x for numerical and analytical water quality models[3,4]. D_x is the most predominant factor influencing the pollutant concentration at the downstream of the point of

[1]School of Environment, College of Engineering, University of Tehran, Tehran 1417853111, Iran. [2]Faculty of Governance, University of Tehran, Tehran 1439814151, Iran. [3]Department of Geospatial Information Systems, College of Engineering, University of Tehran, Tehran 1439957131, Iran. [4]School of Civil Engineering, College of Engineering, University of Tehran, Tehran 1417613131, Iran. [5]College of Hydrology and Water Resources, Hohai University, Nanjing 210098, China. [6]Department of Civil and Environmental Engineering, College of Engineering, Chung-Ang University, Seoul 06974, Korea. [7]Civil and Environmental Engineering and the National Water Center, United Arab Emirates University, Al Ain 15551, Abu Dhabi, United Arab Emirates. [8]Department of Civil and Environmental Engineering and Water Resources Research Center, University of Hawaii at Manoa, Honolulu, HI 96822, USA. [9]School of Engineering, University of Warwick, Coventry CV4 7AL, UK. [10]These authors contributed equally: Behzad Ghiasi, Roohollah Noori and Hossein Sheikhian. ✉email: noor@ut.ac.ir; m.hamouda@uaeu.ac.ae

accidental pollution[5–8]. Starting from the late 1960s, the mechanism of D_x determination in streams was introduced by Fischer[9]. Fischer[10] proposed an analytical formula to estimate D_x that required detailed knowledge of the flow-geometric conditions of the system under study.

Given that the flow-geometric data for streams, especially in large meandrous channels, are highly variable in temporal and spatial scales, such data are not readily measured and available. Also, the complex numerical procedures required to solve Fischer[10] equation, have led to introduction of several simplifications to determine D_x. Hence, the estimations of D_x from the simplified models can largely deviate from the field-based estimated measurements[11,12]. These simplifications are mainly exclusion of some variables which are difficult to access such as flow-geometric irregularities that influence dispersion mechanism in streams. Although in many cases the impact of the excluded variables is somewhat embedded in other variables used in D_x estimation models, they do not fully represent the complex interactions between the absent variables and D_x. For example, friction term (i.e., rate of flow velocity to shear velocity – U/U^*), as a readily accessible input for D_x estimation models, to some extend can represent the impact of lateral and vertical irregularities in streams that affect the rate of dispersion[13]. However, these irregularities produce shear flows and secondary currents that can alternate the D_x. Simultaneously, the former causes an increase in D_x whilst the latter decreases D_x[14–18]. The complex interactions between the flow-geometric data and dispersion mechanism prohibit reaching an accurate estimation of D_x in streams whilst some effective variables on dispersion mechanism are excluded (e.g., stream bed shape factor and sinuosity).

In recent decades, and with the advancement in artificial intelligence (AI) models, they became powerful tools to solve complex engineering problems[19–27]. A number of AI-based studies have been conducted to enhance the accuracy of D_x estimation in turbulent flow systems such as natural streams[28–31]. Given that AI techniques are able to map the complex non-linear input–output relationships even when some important information is missing[32], their applications in estimating the D_x have been investigated by several studies[28–31,33–43]. However, complex nature of dispersion mechanism in turbulent flow systems with variations in both spatial and temporal scales, as well as the inevitable simplification assumptions that are needed for the modelling will result in uncertainty of D_x estimation using AI-based models. The uncertainty in the output of hydrological models is largely resulted by factors such as input-data uncertainty, model uncertainty and parameter uncertainty[44–51]. Intensive efforts have been made to investigate the uncertainty of physics-based hydrological models, which led to good understanding of the different sources of uncertainty and their quantification approaches in hydrological models[44–51]. However, there still remains a significant need to understand and quantify the uncertainty associated with AI-based hydrological models, especially for water quality modelling. In river water quality modelling, the majority of existing AI-based studies are conducted to find the best point estimation, without much attention towards the uncertainty quantification of the model predictions. AI-based models, as data-driven techniques, have not been elaborated to consider the physical mechanisms of the objective parameter under study. In contrast with the physics-based models, AI-based models discover and learn the underlying physical mechanisms that govern water quality parameters using a training process[38,41,42]. The performance of training procedure depends on the sampling patterns selected to tune the AI-based model. Therefore, given that the predictions of AI-based models are highly impacted by the data used for training, any changes in the selected training data can impose large uncertainty in the model output. In a study conducted by Noori et al.[42], they reported that although the AI techniques outperform empirical-based models for estimation of D_x, their predictions are still subject to uncertainty induced by changes in their training patterns. The inaccuracy in estimation of the D_x using AI models can limit water quality assessment and design of appropriate measures to improve the water quality of aquatic flows. Hence, developing a robust methodological framework to quantify the prediction uncertainty of the D_x from AI-based models is essential for developing appropriate AI-based water quality models.

Granular computing (GrC) model is a highly efficient AI-based model which has recently shown an excellent potential to solve complex engineering problems[52–56]. GrC model is a novel tool capable of applying the granules in the process of nonlinear problem solving[52]. In the GrC model, the natural rules between the data are extracted by means of the rule mining algorithm, operating on a set of information arranged as information table. The granule measures involved in the process of information mining, has made GrC as a powerful tool to map a set of inputs to a set of outputs in different fields of science and engineering[52,53]. However, similar to other AI models, the GrC performance can be adversely influenced by the selection of training patterns. Therefore, the effects of changes in training patterns on the performance of GrC model should be investigated, to understand and quantify the degree of uncertainty in the model's prediction of D_x in water quality assessments. Previous studies which examined the application of GrC model for D_x estimation in natural streams did not investigate the prediction uncertainty introduced by the model training patterns[39,43]. In this study, we first coupled an artificial neural network (ANN) with rules information in the GrC (GrC-ANN) to improve the GrC model's performance. Encoding the given information used in the GrC into a feed forward multi-layer structure, i.e. ANN, enhances the GrC model to use all information available in the dataset to decide about different presented patterns. Then, a D_x predictive model was developed using GrC-ANN modelling technique. Finally, a comprehensive uncertainty analysis method was proposed to compare the accuracy of D_x predicted by the GrC-ANN with other AI-based D_x models in the literature. Our proposed method quantifies the GrC-ANN prediction's uncertainty based on the model response to change in the selected training patterns using a modified bootstrap method[12].

Methods

Longitudinal dispersion.

Non-reactive pollutant mixing in aquatic systems is a complex three-dimensional (3-D) flow process, consists of molecular and turbulent diffusion, and shear dispersion (referred to as "dispersion") mechanisms. The dispersion is the net trace of velocity shear over the flow width and depth, and the turbulent mixing[11]. In the natural streams, which are specifically much longer than width or depth of the flow, the pollutants become well-mixed in the vertical and transverse directions rather than the longitudinal

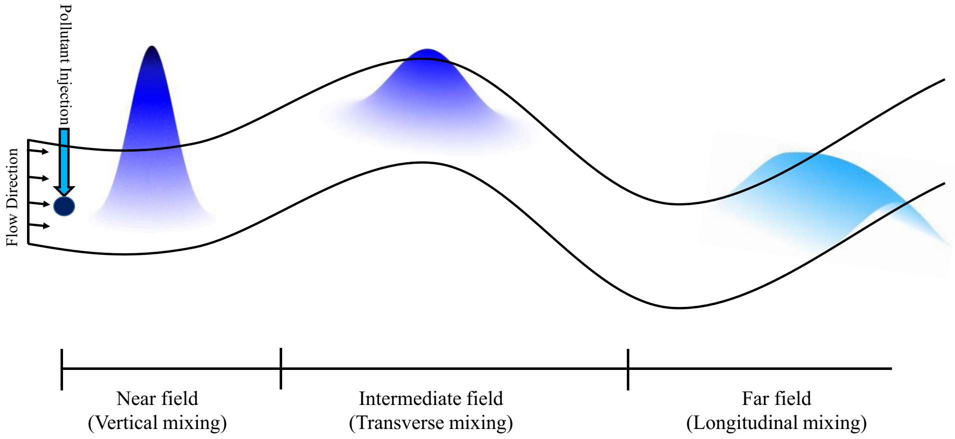


Figure 1. Schematic of concentration profiles and pollutant mixing in streams. (adapted from Kilpatrick and Wilson[59]).

mixing (Fig. 1). Therefore, pollutant fate and transport in streams is usually studied by the application of 1-D mixing model quantified by the advection–dispersion equation as follows[57,58]:

$$\frac{\partial C}{\partial t} + U\frac{\partial C}{\partial x} = D_x\frac{\partial^2 C}{\partial x^2}. \tag{1}$$

In Eq. (1), C and U, are the averaged cross-sectional concentration and averaged longitudinal velocity, respectively, t denotes time and x is the longitudinal coordinate in the stream-wise direction.

Ideally, vertical and (transverse) dispersion in streams takes place close to (in intermediate fields from) the pollutant discharge location, whilst the longitudinal dispersion occurs far from the pollutant discharge point, where solute become readily well-mixed in both vertical and transverse directions (Fig. 1). In streams, the longitudinal dispersion usually varies form 10^{-1} to 10^{7} m^2/s[10,13,60,61] and the diffusion coefficient ranges from 10^{-9} (molecular) to 10^{-2} m^2/s (turbulent)[5]. Therefore, dispersion is the dominant mechanism of mixing process, by several orders of magnitude[62], highlighting the necessity of developing robust methodological approach to quantify the dispersion and mixing coefficient in the streams.

D_x parametrization.

Pioneering work on quantification of dispersion mechanism in pipes date back to Taylor's studies[63,64]. Thereafter, Taylor's approach was used for quantifying dispersion in streams with the assumption of no limits for the width of the channel by Elder[65]. However, the Elder's formula underestimates the dispersion in natural streams, as it does not consider the influence of the lateral velocity shear[10,66]. In streams, the lateral velocity shear mechanism plays a more dominant role in determining the mixing compared to the vertical shear. On this basis, Fischer[9] derived an analytical formula for determining D_x as:

$$D_x = \left\{ \int_0^W h(y)\, \acute{u}\,(y) \int_0^y \left[1/\varepsilon_t(y) h(y)\right] \int_0^y h(y)\, \acute{u}\,(y) dydydy \right\} / A, \tag{2}$$

where, W denotes the local flow width, x is the longitudinal coordinate, y is the lateral coordinate, $\acute{u}\,(y)$. is the local velocity deviation, $h(y)$ represents the local flow depth, $\varepsilon_t\,(y)$ is local lateral mixing coefficient, and A represents the local flow cross-sectional area.

In Eq. (2), the flow is supposed to be 1-D, i.e., the pollutant is well-mixed in both vertical and lateral directions, a condition that is rarely satisfied in turbulent flow systems such as large meandrous streams and even in laboratory flumes, due to existence of secondary currents[67]. Fischer[9] equation has been derived based on the assumption that the dispersion is controlled by lateral shear rather than the vertical shear, a condition that may not be well-satisfied for the narrow and deep rivers where the aspect ratio (i.e., river flow width to depth – W/H) is small[5]. These drawbacks of Eq. (2) lead to inaccurate estimation of D_x compared to those values determined from tracer measurements. The deviation between D_x values estimated by Eq. (2) and those true values is maximum for the case of non-uniform flow in real meandrous streams, albeit Fischer[9] model can well approximate the dispersion for the case of uniform flows[68]. In addition to the inherent drawbacks in practical application of the Eq. (2), it also requires detailed information on the geometrical properties (i.e. cross-section, bathymetry) of stream, as well as the lateral flow velocity profiles. Collecting such information is rather costly and time consuming, and often requires very detailed flow measurements which are not readily available. Therefore, practical application of Fischer[9] model is limited.

To address the difficulties in using Eq. (2), Fischer[69] suggested a simplified empirical equation that correlates D_x with pertinent dimensionless variables of W/H and U/U^* as follows:

$$\frac{D_x}{HU^*} = a\left(\frac{W}{H}\right)^b\left(\frac{U}{U^*}\right)^c. \tag{3}$$

Fischer[69] modified empirical formula for determining the dispersion coefficient (Eq. 3), has been widely used and validated by other researchers[11–13,28–31,70] and rely on the parameters which can be practically determined for natural streams.

Data collection. This study aims to estimate D_x in streams using GrC-ANN model. In this regard, a global tracer database consisting of 503 observations from natural streams and laboratory flumes was used to develop the model and validate the performance of the proposed GrC-ANN model. This database was compiled by Riahi-Madvar et al.[71], and include data on the friction term, aspect ratio, and with D_x ranging between ~ 0.00 to ~ 1800 m^2/s. Although the database used in this study is more comprehensive compared to other studies on D_x estimation, it does not fully include extreme high values of D_x[12]. The reported D_x values in the literature are within the range of near to zero (in the laboratory flumes) to extreme high value of 6800 m^2/s in large and irregular-shaped rivers[72]. The maximum D_x used in this study is ~ 1800 m^2/s, which is related to dispersion in natural streams with irregular hydraulic-geometric characteristics, and dispersion values greater than what is used in this study are extremely rare in environmental hydraulics problems. Therefore, the extremely high D_x values were excluded from the database as outliers, given that they significantly impact the statistical analyses[12]. However, D_x/HU^* parameter in the database used here has a non-normal distribution as described by Noori et al.[12]. Using a preliminary investigation, it was found that no significant difference exists between the GrC-ANN model outputs with the normalized and raw D_x/HU^* data. Therefore, the raw data was considered for further investigations in this study.

GrC-ANN model development. In-depth description of the ANN, GrC and GrC-ANN approaches for D_x modelling in streams and the model development procedures are given in Noori et al.[39] and Ghiasi et al.[43], respectively. Further detailed information about these models documented in the literature[52–55,73–75]. Hence, we shortened the descriptions of GrC-ANN model developed in this study.

- **GrC model**

  Granular Computing models are superset of the rough set theory, interval computations and the theory of fuzzy information granulation[52]. GrC model is a data processing method based on multiple levels of data granularity. In this method, the whole dataset is divided into granules and clusters (or subsets), which categorizes individual elements of the whole dataset based on the existing similarity between objects to put them in different granules. Then, a set of rules is extracted over concepts ϕ and Ψ in the form of IF–THEN: "If an objective satisfies ϕ, THEN the object satisfies Ψ". Here, concepts ϕ and Ψ are a set of attribute-values for a set of objects and the assigned output value, respectively. In the process of rule extraction, GrC algorithm forms all the possible granules to extract every relation between the patterns, i.e. extracted rules, regardless of their importance or accuracy. Following rules extraction procedure, the algorithm applies statistical measures on granules formed in order to select the best set of possible rules, i.e. pruned rules, to form the regression rule set[52–55].

  Generality (G), absolute support (AS), coverage (CV), and conditional entropy (CE) are the statistical measures used by the GrC to extract the rules. The generality of concept ϕ, i.e. $G(\phi)$, displays the relative size of constructive granule of this concept, defined by Eq. (4)[76]:

$$G(\phi) = \frac{|m(\phi)|}{|U|}, \tag{4}$$

  where $|m(\phi)|$ is the size of the granule and $|U|$ is the size of the entire domain. $G(\phi)$ varies between 0 and 1. Higher values of generality describe the rule as a more common concept, which is more probable to occur. On the other hand, high $G(\phi)$ can bias the model towards the patterns observed during the training process.

  AS, as the conditional probability in the case that a randomly selected object satisfies both ϕ and Ψ, can be obtained from Eq. (5) and describes the strength of a rule in assigning similar outputs to a set of input values[73]. $AS = 1$, if and only if $m(\phi) \subseteq m(\Psi)$.

$$AS(\phi \to \Psi) = \frac{|m(\phi \Lambda \Psi)|}{|m(\phi)|}. \tag{5}$$

  CE, represented by $H(\Psi \mid \phi)$, reveals the uncertainty of formula ϕ based on formula Ψ and is defined by Eq. (6)[73]. CE ensures the model reliability and robustness, by filtering out the rules that are providing information which is not supported by other rules in the rule set, even if these rules have misleading acceptable values for other statistical measures.

$$H(\Psi \mid \phi) = -\sum_{i=1}^{n} p(\Psi_i \mid \phi)\log\big(p(\Psi_i \mid \phi)\big), \tag{6}$$

  where, $p(\Psi_i \mid \phi) = |m(\phi \cap \Psi_i)|/|m(\Psi)|$.

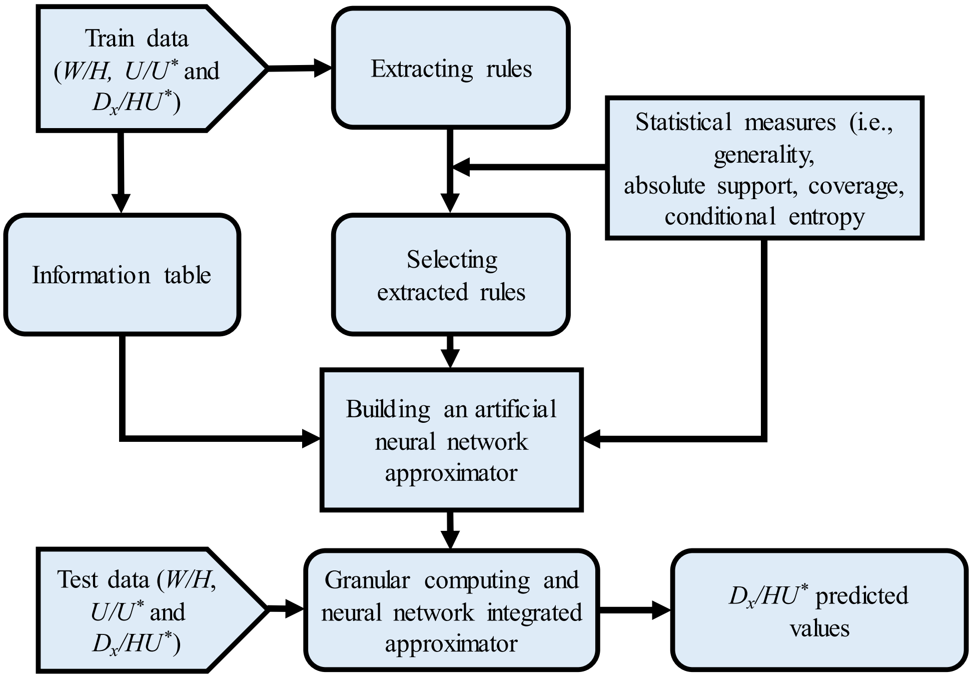

Train data (W/H, U/U* and Dx/HU*)
Extracting rules
Statistical measures (i.e., generality, absolute support, coverage, conditional entropy
Information table
Selecting extracted rules
Building an artificial neural network approximator
Test data (W/H, U/U* and Dx/HU*)
Granular computing and neural network integrated approximator
Dx/HU* predicted values

Figure 2. Procedure of integrating GrC and ANN for determining longitudinal dispersion coefficient.

CV denotes the conditional probability of a randomly selected object to satisfy ϕ, while satisfying Ψ[73]. This parameter shows the strength of a rule in predicting accurate output values if different training patterns are provided to the model.

$$CV(\phi \to \Psi) = \frac{|m(\phi \cap \Psi)|}{|m(\Psi)|}. \tag{7}$$

In this study, GrC extracts the rules from the global tracer database consisting of 503 observations from natural streams and laboratory flumes based on the *CE* and *AS* statistical measures, so that the rules with the minimum *CE* value and the maximum *AS* are extracted from the database. To form a granular decision tree, the priority of rules in the tree is determined based on the *G* and *CV*.

- **ANN model**

  An ANN consists of a set of neurons, as the smallest computational units of the model, organized in different layers joint by connection weights. The first and last layers are the input and output layers of the network, respectively. The layers in the middle of the network are hidden and contain computing neurons. To construct an ANN for a predictive modelling purpose, training data are introduced to the network. Then, the network starts the learning process by determining connection weights and biases based upon the resulting error at the output nodes[77]. Upon obtaining the connection weights and biases, the network is ready to do a classification or regression task.

- **GrC-ANN model**

  A basic GrC model has two major deficiencies. First, it prioritizes rules based on their obtained parameters and uses the first rule satisfied by the input data to define its output. Second, it cannot make use of information provided in the rule set and makes a prediction by only using one rule[73–75]. Hence, to compensate for these deficiencies, the GrC-ANN model proposed in this study uses an integration of GrC rule generation algorithm and ANN model (Fig. 2). The GrC-ANN approach allows the model to use the mentioned rule quality parameters (i.e. *G*, *AS*, *CV*, and *CE*) to construct the approximator structure, instead of common time-consuming iterative learning procedure used by ANN model[48]. Given the input patterns, the GrC-ANN model tunes the network by re-forming the granules and applying statistical measurements performed by the GrC approach. Re-forming the granules also re-forms the rules, which results in different number of rules and different statistical measurements. *CE* plays an important role in tuning the model. Keeping *CE* close to zero filters out inconsistent rules by removing them or giving them less importance. GrC-ANN tries to minimize the number of rules by minimization of *CE* and maximization of *G*, *AS* and *CV*[52,53].

The GrC-ANN structure proposed in this paper, similar to the conventional neural networks, comprises of layers including the input layer, two computing layers, and the output (aggregation) layer (Fig. 3). The layers within the proposed GrC-ANN model are customized to ensure robust predictions of D_x. The number of nodes in the input layer are set equal to attributes of the data records (i.e., *W/H*, *U/U**, and D_x/HU^*). Computing layers are comprised of two inner-connected layers including pattern layer and rule firing layer. The computing layers receive values that are valid according to the criteria determined in the input layer. Computing layers'

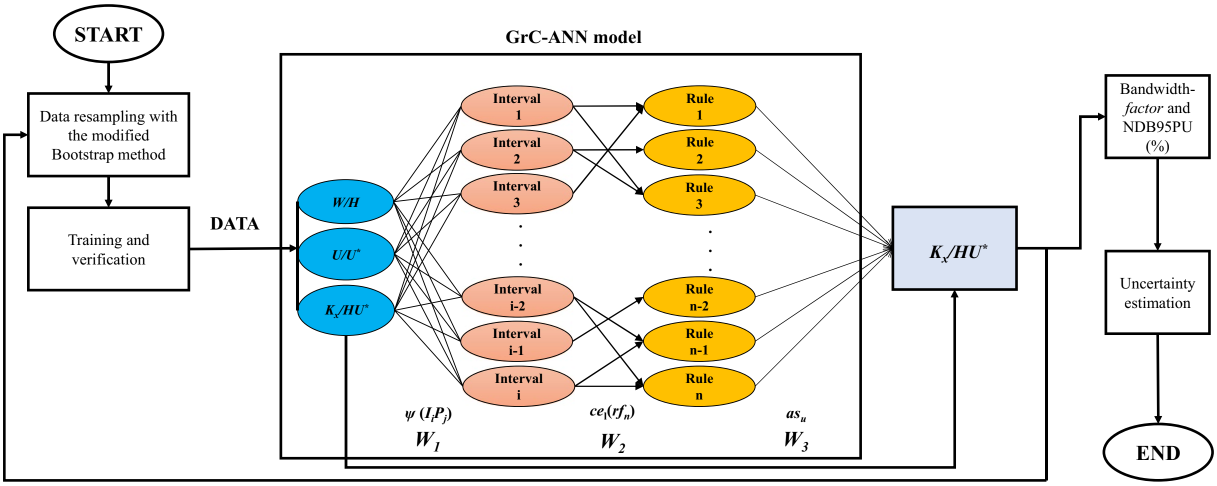


Figure 3. Schematic of the GrC-ANN model structure developed for this study.

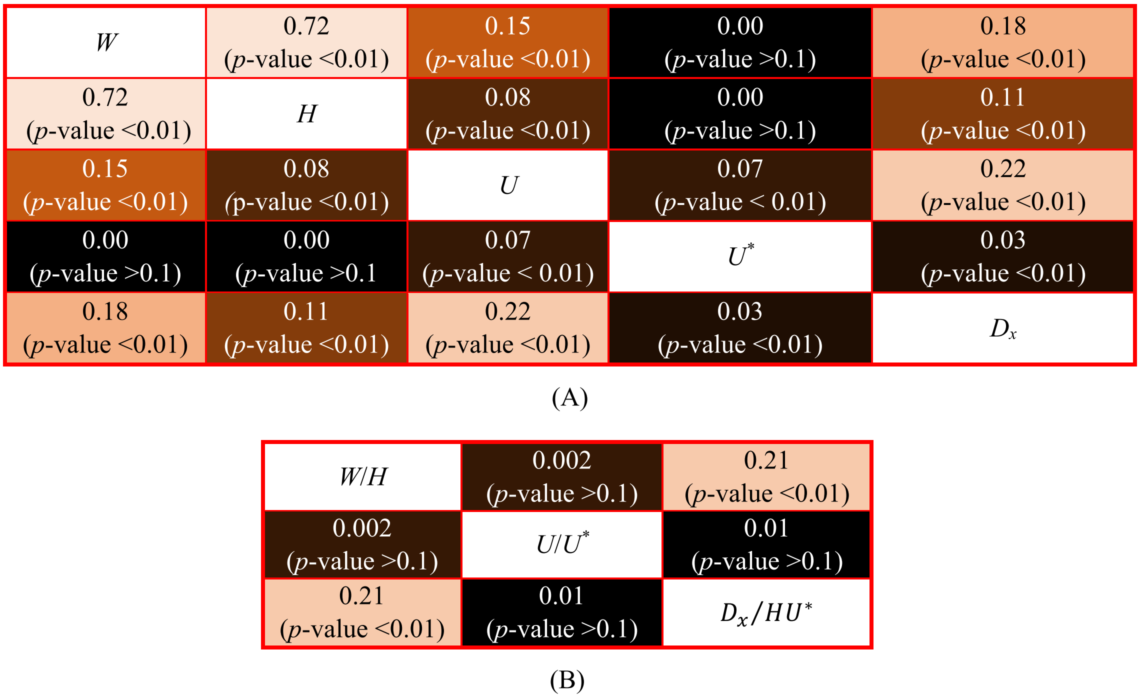
characteristics are fully data driven. Pattern layer nodes act as transformers, normalizing quantized valid values of the criteria in the input layer as the rule firing nodes expect. Rule firing nodes use the data provided from the measured and selected rules to aggregate the received values, turning them into predictions. The third layer contains the set of qualified extracted rules by GrC-ANN and embeds the classification rules. The aggregation layer assigns an output value to the input pattern of the data. The connection weights of the rule-firing layer and the aggregation layer are given by the statistical measure of absolute support provided by the corresponding rule to its output value, to consider the accuracy of the rules in determining that output value[43,52,53].

The proposed GrC-ANN approach benefits from some advantages that are absent in both GrC and ANN models. Utilizing tangible information obtained from the rule measures in the form of neurons, layers and connection weights improve the transparency of the constructed model[43]. Since the given information are encoded in a feed forward multi-layer structure, similar to ANNs, the GrC-ANN will be able to use all information available in the dataset to decide about each presented data pattern, which is an improvement to rule-based classifiers, such as GrC[52]. Replacing the learning part of an ANN with the information from rule quality measures ensures that no connections or nodes are remained without a transparent description. This is an improvement to the conventional ANNs which contain hidden neurons and obtain their connection weights and biases by learning through a black-box learning algorithm[43]. A conventional ANN provides results which is influenced by initial weights generated in a random manner, yielding to different results from the same set of training information, lacking the ability to describe them. GrC-ANN provides a robust network and can be manipulated by defining rule measure thresholds. Overall, these advantages improve the accuracy of the proposed GrC-ANN predictions compared to conventional GrC and ANN models[43,52]. Although GrC-ANN model reduces the computational time needed for model training by removing the learning procedure in the ANN model, it requires more computational cost than ANN model due to the procedure of extracting high-quality classification rules. In general, the computational cost for the GrC-ANN model is in the order of: $\mathrm{O}\left(n^2 \times p_1 \times a \times m + \times p_2 \times n \times a \times (l \times r)\right)$, aggregating training and verification time, where n is the number of iterations, a is the number of attributes for the patterns, m is the number of GrC measure parameters, r denotes the number of rules used in prediction, l is the number of layers in the network, p_1 and p_2 are the number of input patterns for training and verification, respectively[53,73–75].

Uncertainty quantification.

Similar to other data-driven models, the GrC-ANN model minimizes the error function based on the data fed with the aid of a supervised algorithm throughout the training process[43]. Hence, model training plays a vital role in quantification of the GrC-ANN model's uncertainty caused by different tuning sets. In this study, the GrC-ANN model was tuned to map the input parameters, i.e. W/H and U/U^*, to the target D_x/HU^*, based on finite training patterns resampled from 503 observations of the global tracer database. Probabilistically, each training pattern used for tuning the GrC-ANN model is different from others resampled from the global database. Thus, each training pattern could produce different set of GrC-ANN parameters, and predictive outputs for the estimation of D_x/HU^*.

The modified bootstrap method suggested by Noori et al.[12] was used to resample distinct training patterns for tuning the GrC-ANN model for D_x/HU^* predictions. This method ensures that the chosen training patterns are fully representative of the statistical characteristics of the 503 tracer experiments of the global database used in this study. This is particularly important since the global database used in this study rarely has large D_x instances[12], denoting that these large dispersion values are likely to be under-represented in the training patterns chosen by the conventional bootstrap technique. This issue can result in poor training of the GrC-ANN model and consequently increases the model's uncertainty in prediction of D_x/HU^*. Detailed description of the bootstrap method is given by Efron and Tibshirani[78], while the modified the bootstrap method adopted in this study is described by Noori et al.[12].



W	0.72 (p-value <0.01)	0.15 (p-value <0.01)	0.00 (p-value >0.1)	0.18 (p-value <0.01)
0.72 (p-value <0.01)	H	0.08 (p-value <0.01)	0.00 (p-value >0.1)	0.11 (p-value <0.01)
0.15 (p-value <0.01)	0.08 (p-value <0.01)	U	0.07 (p-value < 0.01)	0.22 (p-value <0.01)
0.00 (p-value >0.1)	0.00 (p-value >0.1	0.07 (p-value < 0.01)	U^*	0.03 (p-value <0.01)
0.18 (p-value <0.01)	0.11 (p-value <0.01)	0.22 (p-value <0.01)	0.03 (p-value <0.01)	D_x

(A)

W/H	0.002 (p-value >0.1)	0.21 (p-value <0.01)
0.002 (p-value >0.1)	U/U^*	0.01 (p-value >0.1)
0.21 (p-value <0.01)	0.01 (p-value >0.1)	D_x/HU^*

(B)

Figure 4. The correlation coefficient plots of (**A**) W, H, U, U^*, and D_x, (**B**) W/H, U/U^* and D_x/HU^*.

Different outputs of the D_x/HU^* GrC-ANN model in the verification stage, i.e. due to the change in the training patterns, were used as a measure of the model's uncertainty[79]. An interval band of the GrC-ANN estimations of D_x/HU^* was computed, with a level of significance of 95%. Then, two measures were introduced to assess the D_x/HU^* prediction variations in the different responses of the GrC-ANN model in verification stage including bandwidth-*factor* and the number of bracketed D_x/HU^* data using 95% of predicted uncertainties (NBD95PU) as shown in Eqs. (8) and (9), respectively[80]. Given these two measures, the uncertainty in estimation of the D_x/HU^* GrC-ANN model in verification stage was quantified.

$$\text{bandwidth}-\textit{factor} = \left\{ \left(\frac{1}{n}\right) \sum_{i=1}^{n} (X_U - X_L) \right\} / \sigma_x, \tag{8}$$

$$\text{NBD95PU}(\%) = (1/n)\text{count}\{Q|(X_L \le Q \le X_U)\}, \tag{9}$$

where σ_x is the standard deviation of the target D_x/HU^*, and X_U and X_L are the maximum and minimum of the estimated D_x/HU^* for each training pattern, respectively.

Figure 3 illustrates a detailed description of the model development and uncertainty quantification process proposed for this study.

Results and discussion

Tuned GrC-ANN models. The correlation amongst the input parameters, i.e. W, H, U, U^*, and D_x is shown in Fig. 4A. The correlation coefficients for the model variables in dimensionless format, i.e. W/H, U/U^* and D_x/HU^* and the corresponding statistical significance level are illustrated in Fig. 4B. In dimensional form, D_x/HU^* is more correlated with the geometrical configuration W/H of the stream (correlation coefficient = 0.21, p-value < 0.1) than the flow characteristic U/U^* (correlation coefficient = 0.002, p-value > 0.1), confirming the results reported by Noori et al.[12].

To examine the GrC-ANN model, the database with 503 observations from natural streams and laboratory flumes were scaled between 0 and 1. 40 data instances were selected from the global tracer database for the model verification. Then, 100 distinct training patterns were randomly resampled from the remaining database, i.e. 463 observations, with replacement to tune 100 different D_x/HU^* GrC-ANN models. Each training pattern consists of 80 data, and the 40 pre-assigned verification data. The model inputs include, aspect ratio and friction term, and dimensionless target D_x/HU^* were clustered based on their indiscernibility in the given attributes. To form the final rule network, GrC-based rule extraction algorithm was used to select the best granules of information by considering the *CE*, *AS*, *G*, and *CV* measures computed for each rule. In this regard, *AS* and *CE* indices were employed to extract the set of possible valid rules by considering minimum and maximum threshold values of 0.75 and 0.5, respectively, in accordance to similar studies in the literature[39,81,82]. At this stage, if a rule caused redundancy in the rule set, it was considered as an active granule and was replaced with a granule that had more consistency in the set of rules. Using the proposed methodology led to extraction of a range of rules, varied from 76 to 234, for tuning the GrC models based on the training patterns (Fig. 5A).

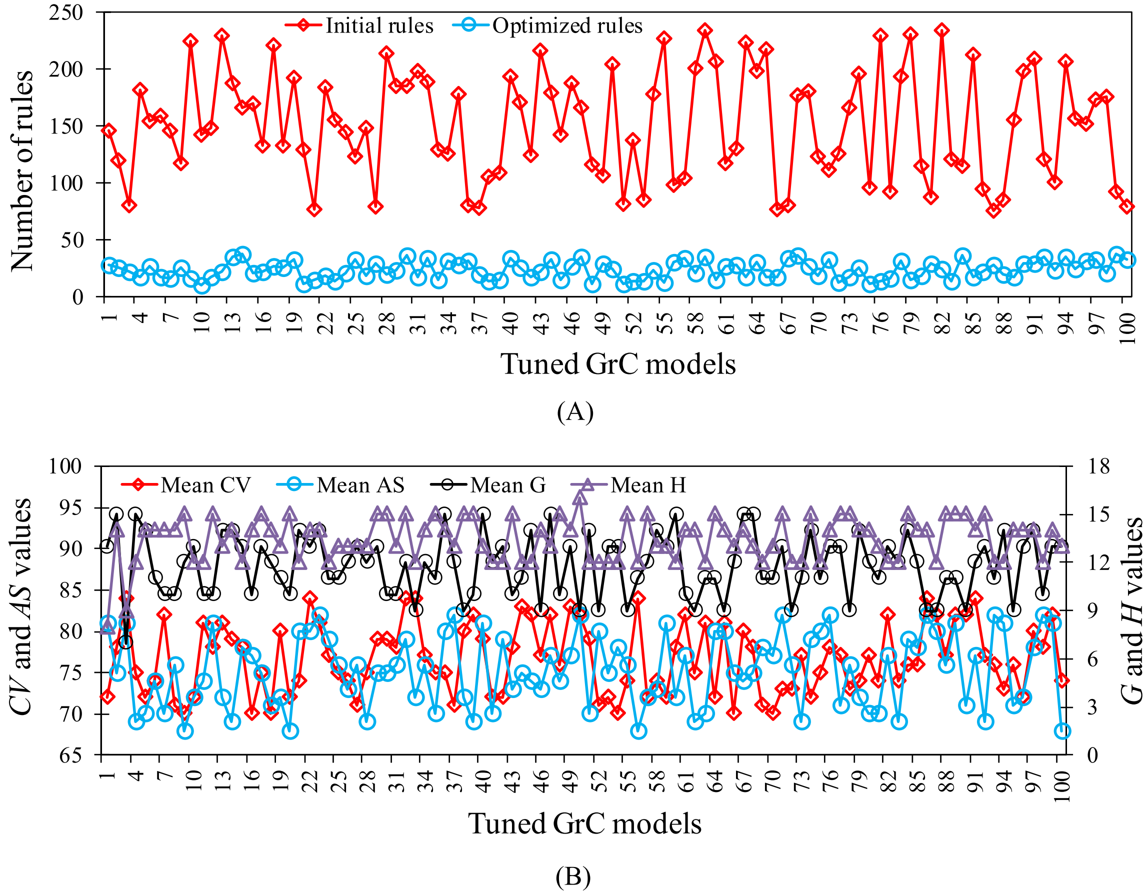


Figure 5. (**A**) Number of initial and optimized rules, and (**B**) the mean values of quality indices for the final rules for each tuned GrC models.

In the next step, the *CV* and *G* indices were applied to prioritize the rules that construct the final rule sets. For the models tuned based on the training patterns, the optimized rules varied from 10 to 38 (Fig. 5A). The mean values of quality indices for the final rules selected for each tuned model are illustrated in Fig. 5B. According to Fig. 5B, the *G* values ranged between 0 and 0.4, indicating the rules' generality does not pertain to big values of *G*, confirming the results of previous GrC modelling studies[39,74]. The *CV* varied between 0 and 1, pertaining to the numbers of extracted rules by each class and the dataset covered by each rule, following Yao and Yao[74] findings.

The 100 optimized rule sets computed correspond to one hundred distinct training patterns, which are then fed to the GrC-ANN modelling structure. In this regard, the rule quality indices were embedded into an ANN structure instead of initial weights, forming a GrC-ANN model corresponding to each optimized rule set. The best network structures describing the relations between the inputs (W/H and U/U^*) and the output (D_x/HU^*) data were determined based on the quality index of root mean square error (RMSE) for each GrC-ANN model tuned by the distinct training patterns (Fig. 6A). Analysis of the results show the RMSE values for the tuned D_x/HU^* GrC-ANN models, in training and verification stages varied from 1251 to 2142 and 966 to 3826, respectively (Fig. 6A).

Figure 7 shows the difference between the true (field-estimated) D_x/HU^* values and those predicted by each tuned D_x/HU^* GrC-ANN model. The minimum (i.e., – 10,934) and the maximum (i.e., 7471) errors were produced by D_x/HU^* GrC-ANN models #42 and #100, respectively. In general, the GrC-ANN models overestimate the D_x/HU^* values for approximately 86% of the observations (Fig. 6B). Similar overestimation of D_x was reported by Etemad-Shahidi and Taghipour[83] for the D_x models proposed by Liu[61], Seo and Cheong[13], Deng et al.[57], and Sahay and Dutta[84]. In this study, the overestimation of D_x/HU^* could be associated with the RMSE values used as the objective function in the GrC-ANN model. RMSE is a scale-dependent parameter and could lead the model to predict values with lower relative error for large D_x values that rarely exist in the database. In addition, we defined a constraint for the GrC-ANN model to filter out the modeling result for the negative values, which are likely to contribute to the overestimation for small D_x/HU^* values that are the dominant feature of the database. However, using the overestimated D_x/HU^* values in 1-D ADE models give a lower maximum concentration rate for those locations which are far from the pollutant injection point[12]. Therefore, the tuned D_x/HU^* GrC-ANN model must be used with caution in hydro-environmental studies such as outfall design, and risk assessment studies for accidental hazardous pollution.

Comparative analysis of the tuned GrC-ANN models developed in this study, and other AI models including model tree (MTree), gene-expression programming (GEP), evolutionary polynomial regression (EPR), support

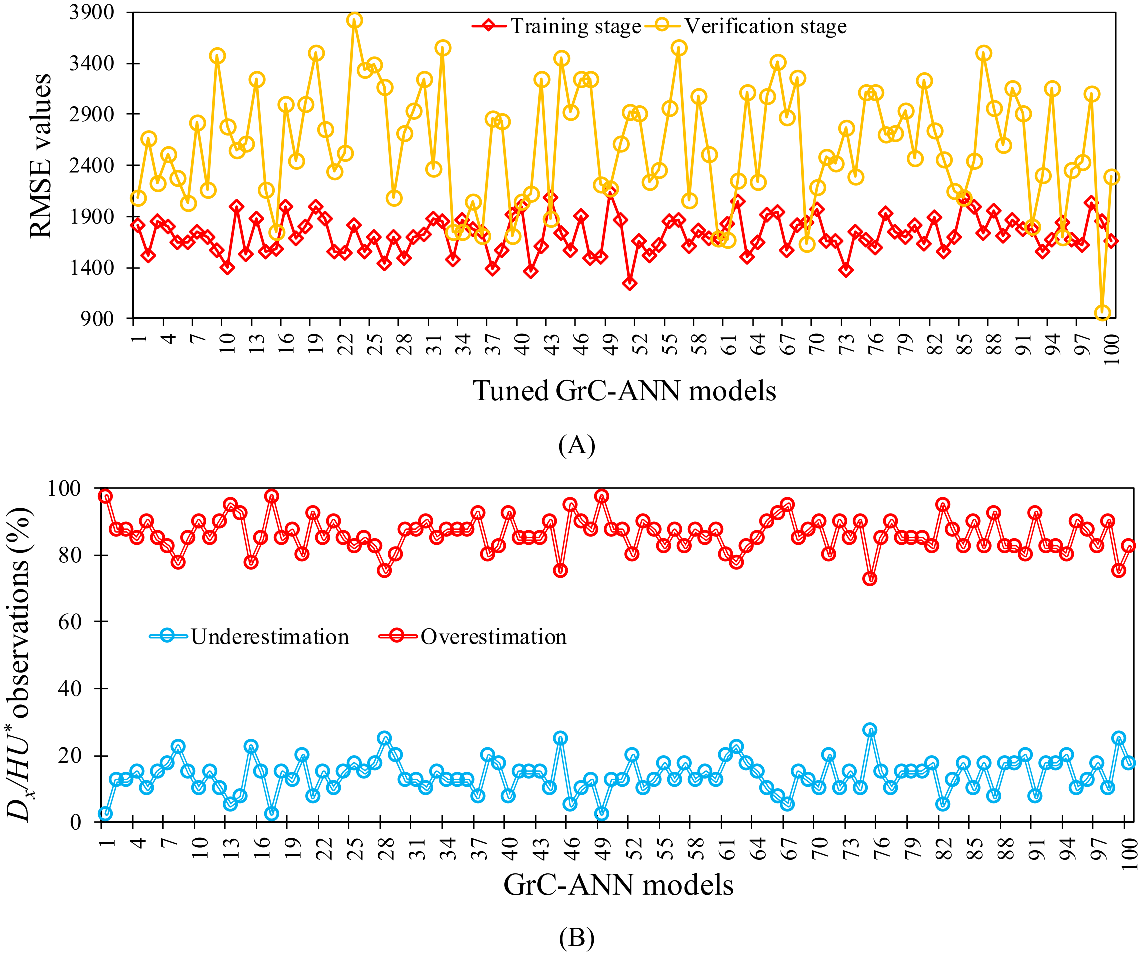


Figure 6. (**A**) Root mean square error (RMSE) values calculated for the tuned D_x/HU^* GrC-ANN models in training and verification stages, and (**B**) D_x/HU^* observations (%) with underestimation and overestimation in GrC-ANN models tuned by the distinct training patterns.

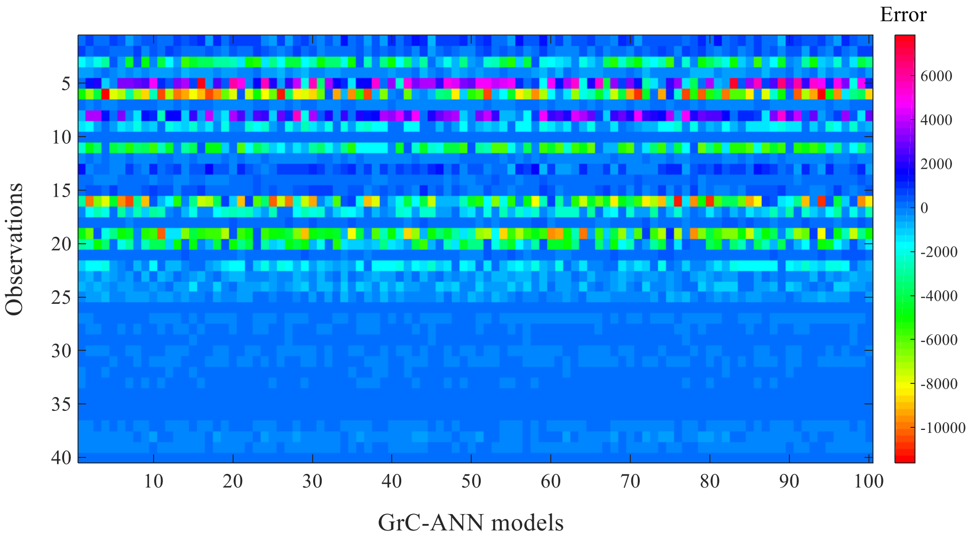


Figure 7. Difference between the true D_x/HU^* values and those predicted using GrC-ANN models tuned by the distinct training patterns during the verification stage of the model.

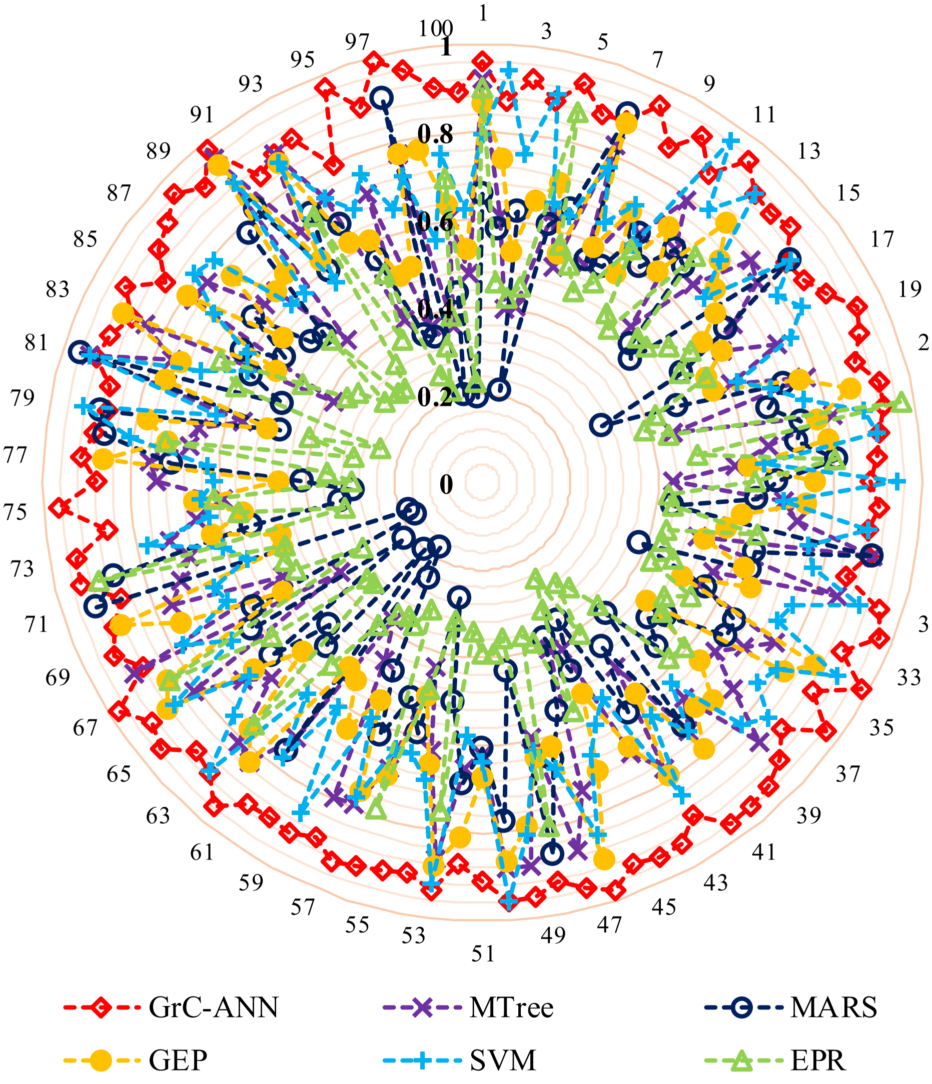


Figure 8. Determination of coefficient (R^2) values calculated for D_x/HU^* prediction during the verification step of the tuned GrC-ANN models developed in this study, and those reported for model tree (MTree), gene-expression programming (GEP), evolutionary polynomial regression (EPR), support vector machine (SVM) and multivariate adaptive regression splines (MARS) by Najafzadeh et al.[85].

vector machine (SVM1), and multivariate adaptive regression splines (MARS), developed by Najafzadeh et al.[85], highlights that the proposed GrC-ANN models are capable of better and more robust approximation of longitudinal dispersion (D_x/HU^*) in streams (Fig. 8). Previous studies also confirmed the performance superiority of GrC compared to ANN and adaptive neuro fuzzy inference system (ANFIS) developed for D_x/HU^* predictions[43]. As shown in Fig. 8, the determination coefficient (R^2) values determined for the GrC-ANN models in verification stage, are much larger than those reported for ERP and MARS models. However, the computational cost of GrC-ANN model is more than that for ANN models. In this study, the computational time of GrC-ANN models was approximately 1.8 to 2.6 times greater than that for the ANN models.

GrC-ANN uncertainty. The D_x/HU^* values estimated during the verification stage by the 100 GrC-ANN models tuned under distinct training patterns were used to measure the model uncertainty. In this regard, prediction intervals corresponding to each D_x/HU^* observation was computed by considering the level of significance of 95% (Fig. 9). These prediction intervals show the deviation from the true D_x/HU^* values, denoting the uncertainty associated with the GrC-ANN predictions of longitudinal dispersion in streams.

Figure 9 shows that the true D_x/HU^* values are fully located between the lower and upper bands of the uncertainty, concluding the appropriate performance of the GrC-ANN model based on the NDB95PU (%) index. Also the small value of the bandwidth-*factor* (=0.56) indicates the small deviation of the predicted D_x/HU^* values by the GrC-ANN models from the measured values, leading to low uncertainty of the model. Figure 9 shows that the proposed GrC-ANN model has good performance in predicting both large and small D_x/HU^* values with a narrow bandwidth of uncertainty, highlighting the model superiority in predicting the D_x/HU^* compared to other AI models which are suffering from large uncertainty in estimation of D_x[12,42,85].

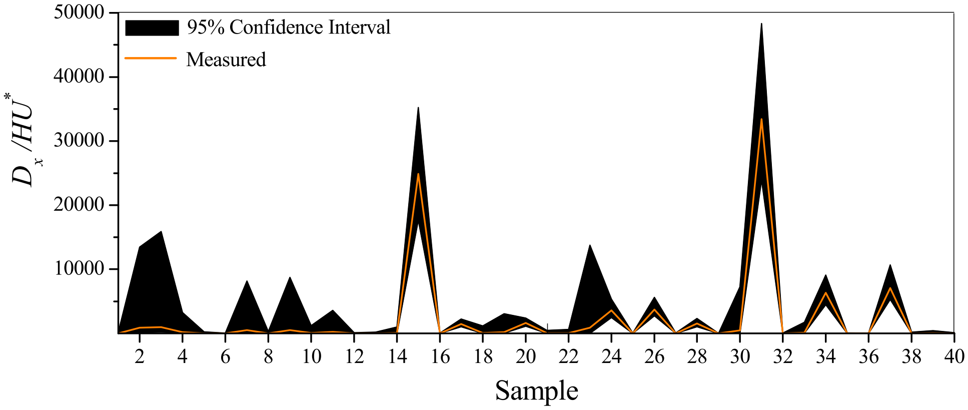


Figure 9. GRC-ANN model uncertainty for estimation of D_x/HU^* in streams.

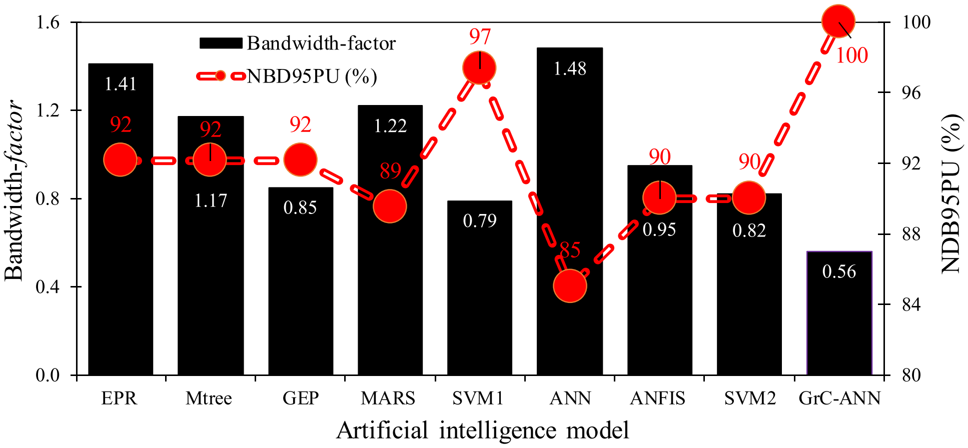


Figure 10. Comparison of the bandwidth-*factor* and the NDB95PU (%) values of the GrC-ANN model developed in this study (i.e., D_x/HU^* GrC-ANN), with ANN and ANFIS models[42], SVM1, GEP, MTree, MARS, and EPR models[85], and SVM2 model[42].

However, neither the GrC-ANN model nor other mathematical and statistical models can fully understand and predict the dispersion processes in real streams. Therefore, the results illustrated in Fig. 9 still contain some degree of uncertainty in the prediction of D_x/HU^* from GrC-ANN model. To compare the uncertainty of the predicted D_x/HU^* from GrC-ANN with other AI models, the bandwidth-*factor* and NDB95PU (%) values computed for these models are illustrated in Fig. 10. This figure shows that D_x/HU^* GrC-ANN model has the smallest bandwidth-*factor* value amongst the nine AI-based models examined in this study. Also, D_x/HU^* GrC-ANN model has the largest NDB95PU (%) value compared to other AI models (i.e., EPR, MTree, GEP, SVM, MARS, ANN, and ANFIS). These measures suggest that the uncertainty in the prediction of D_x/HU^* from GrC-ANN model is far less than those reported for other well-established AI models for the case of pollutant transport in streams.

However, study of the Fig. 9 reveals that despite modified and enhanced training patterns adopted in this study, there remains some uncertainty in the prediction of the D_x/HU^* from GrC-ANN model, which can be considerable at times and leading to a wide confidence interval band for some samples. In fact, in the D_x/HU^* GrC-ANN modelling process, some rules are eliminated due to low criteria values (i.e., *G*, *AS*, *CV*, and *CE*). Therefore, the selected rules, which govern the final prediction of the model, do not fully represent the complex mechanisms of the longitudinal dispersion in streams, leading to inevitable uncertainty in the predictions by GrC-ANN model. In addition, diversity of streams and the irregularities in geometric characteristics and non-linearity of the flow hydrodynamics add to the complexity of the mixing mechanisms in the streams. Therefore, full identification, quantification and inclusion of these intricate natural processes in a mathematical or statistical model is not possible. This is correct even for the non-simplified models for prediction of D_x, i.e. Equation (2), where estimated D_x values are still not in full agreement with those values measured in the field. For example, the minimum error between the estimated and field-measurement of D_x values occurs for the case of a uniform flow, that is usually less than 30%[68]. In the case of non-uniform flow in large meandrous streams with severe irregularities in bathymetry, and spatiotemporal variations in flow hydrodynamics, the estimated D_x using Eq. (2) largely deviates from the true values[11]. The problem of inaccuracy in modelling predictions raises up when using Eq. (3), derived based on simplified assumptions for Eq. (2), and by exclusion of important parameters influencing D_x such as S_f and S_n[5,11,16,86–88]. These excluded parameters are seldom monitored in natural streams due to the difficulties associated with their measurement. Another factor that contribute to the uncertainty in prediction of longitudinal dispersion from GrC-ANN model is the rare presence of very large D_x values in the dataset used in this study. Analysis of the dataset used in this study shows that only around 1% of the 503 global dataset of tracer experiments consists of $D_x > 1000$ m^2/s, whilst the maximum value of D_x in the dataset is around 1800 m^2/s[12]. This absence of very large D_x in the dataset, is leading to uncertainty in the D_x/HU^* predicted by the GrC-ANN model.

Conclusions

Longitudinal dispersion coefficient (D_x) influences the transport and fate of pollutants in streams. Given the high spatiotemporal variability of D_x, previous AI models with single training pattern cannot capture the uncertainty associated with the predictive models for D_x in streams. This study provides rigorous methodological approach to examine and quantify the uncertainty in the prediction of D_x/HU^* from the proposed GrC-ANN model. The detailed analysis of the results highlights that although D_x/HU^* predicted by GrC-ANN model outperforms other AI-based dispersion models, there remains some uncertainty in the predicted D_x from the model which need careful consideration and evaluation. This finding suggests that river water quality assessments and environmental management studies should consider the impacts of uncertainty associated with the D_x estimation on the pollutant concentrations, that could result in detrimental impacts on aquatic biodiversity, and ecosystem function in streams as well as the public health. Enhanced data on the flow hydrodynamics and the geometric features in streams (e.g., stream sinuosity and bed shape factor) for the D_x models can further reduce the uncertainty in estimation of longitudinal dispersion parameter.

Data availability

The data used in this study can be obtained from https://doi.org/10.1007/s11269-018-2139-6.

Received: 31 October 2021; Accepted: 7 March 2022
Published online: 17 March 2022

References

1. Bostanmaneshrad, F. *et al.* Relationship between water quality and macro-scale parameters (land use, erosion, geology, and population density) in the Siminehrood River Basin. *Sci. Total Environ.* **639**, 1588–1600. https://doi.org/10.1016/j.scitotenv.2018.05.244 (2018).
2. Noori, R., Berndtsson, R., Hosseinzadeh, M., Adamowski, J. F. & Abyaneh, M. R. A critical review on the application of the National Sanitation Foundation Water Quality Index. *Environ. Pollut.* **244**, 575–587. https://doi.org/10.1016/j.envpol.2018.10.076 (2019).
3. Ramezani, M., Noori, R., Hooshyaripor, F., Deng, Z. & Sarang, A. Numerical modelling-based comparison of longitudinal dispersion coefficient formulas for solute transport in rivers. *Hydrol. Sci. J.* **64**(7), 808–819. https://doi.org/10.1080/02626667.2019.1605240 (2019).
4. Abolfathi, S., Cook, S., Yeganeh-Bakhtiary, A., Borzooei, S. & Pearson, J. M. Microplastics transport and mixing mechanisms in the nearshore region. *Coast. Eng. Proc.* https://doi.org/10.9753/icce.v36v.papers.63 (2020).
5. Rutherford, J. C. *River Mixing* 347 (Wiley, 1994).
6. Abolfathi, S. & Pearson, J. M. Application of smoothed particle hydrodynamics (SPH) in nearshore mixing: A comparison to laboratory data. *Coast. Eng. Proc.* https://doi.org/10.9753/icce.v35.currents.16 (2017).
7. Cook, S. *et al.* Longitudinal dispersion of microplastics in aquatic flows using fluorometric techniques. *Water Res.* **170**, 115337. https://doi.org/10.1016/j.watres.2019.115337 (2020).
8. Cheme, E. K. & Mazaheri, M. The effect of neglecting spatial variations of the parameters in pollutant transport modeling in rivers. *Environ. Fluid Mech.* **21**(3), 587–603. https://doi.org/10.1007/s10652-021-09787-5 (2021).
9. Fischer, H. B. The mechanics of dispersion in natural streams. *J. Hydraul. Div.* **93**(6), 187–216. https://doi.org/10.1061/JYCEAJ.0001706 (1967).
10. Fischer, H.B. *Methods for Predicting Dispersion Coefficients in Natural Streams: With Applications to Lower Reaches of the Green and Duwamish Rivers, Washington*, vol. 582. (US Government Printing Office, 1968).
11. Deng, Z. Q., Bengtsson, L., Singh, V. P. & Adrian, D. D. Longitudinal dispersion coefficient in single-channel streams. *J. Hydraul. Eng.* **128**(10), 901–916. https://doi.org/10.1061/(ASCE)0733-9429(2002)128:10(901) (2002).
12. Noori, R. *et al.* Reliability of functional forms for calculation of longitudinal dispersion coefficient in rivers. *Sci. Total Environ.* https://doi.org/10.1016/j.scitotenv.2021.148394 (2021).
13. Seo, I. W. & Cheong, T. S. Predicting longitudinal dispersion coefficient in natural streams. *J. Hydraul. Eng.* **124**(1), 25–32. https://doi.org/10.1061/(ASCE)0733-9429(1998)124:1(25) (1998).
14. Nezu, I., Tominaga, A. & Nakagawa, H. Field measurements of secondary currents in straight rivers. *J. Hydraul. Eng.* **119**(5), 598–614. https://doi.org/10.1061/(ASCE)0733-9429(1993)119:5(598) (1993).
15. Deng, Z. Q. & Singh, V. P. Mechanism and conditions for change in channel pattern. *J. Hydraul. Res.* **37**(4), 465–478. https://doi.org/10.1080/00221686.1999.9628263 (1999).
16. Marion, A. & Zaramella, M. Effects of velocity gradients and secondary flow on the dispersion of solutes in a meandering channel. *J. Hydraul. Eng.* **132**(12), 1295–1302. https://doi.org/10.1061/(ASCE)0733-9429(2006)132:12(1295) (2006).
17. Bashitialshaaer, R. *et al.* Sinuosity effects on longitudinal dispersion coefficient. *Int. J. Sustain. Water Environ. Syst.* **2**(2), 77–84 (2011).
18. Nikora, V. & Roy, A. G. Secondary flows in rivers: Theoretical framework, recent advances, and current challenges. *Gravel Bed Rivers Process. Tools Environ.* https://doi.org/10.1002/9781119952497.ch1 (2012).
19. Kişi, Ö. Modeling monthly evaporation using two different neural computing techniques. *Irrig. Sci.* **27**(5), 417–430. https://doi.org/10.1007/s00271-009-0158-z (2009).
20. Khatibi, R., Ghorbani, M. A., Kashani, M. H. & Kisi, O. Comparison of three artificial intelligence techniques for discharge routing. *J. Hydrol.* **403**(3–4), 201–212. https://doi.org/10.1016/j.jhydrol.2011.03.007 (2011).
21. Abolfathi, S., Yeganeh-Bakhtiary, A., Hamze-Ziabari, S. M. & Borzooei, S. Wave runup prediction using M5' model tree algorithm. *Ocean Eng.* **112**, 76–81. https://doi.org/10.1016/j.oceaneng.2015.12.016 (2016).
22. Granata, F., Papirio, S., Esposito, G., Gargano, R. & De Marinis, G. Machine learning algorithms for the forecasting of wastewater quality indicators. *Water* **9**(2), 105. https://doi.org/10.3390/w9020105 (2017).

23. Jaramillo, F. *et al.* On-line estimation of the aerobic phase length for partial nitrification processes in SBR based on features extraction and SVM classification. *Chem. Eng. J.* **331**, 114–123. https://doi.org/10.1016/j.cej.2017.07.185 (2018).
24. Borzooei, S. *et al.* Application of unsupervised learning and process simulation for energy optimization of a WWTP under various weather conditions. *Water Sci. Technol.* **81**(8), 1541–1551. https://doi.org/10.2166/wst.2020.220 (2020).
25. Kamrava, S., Im, J., de Barros, F. P. & Sahimi, M. Estimating dispersion coefficient in flow through heterogeneous porous media by a deep convolutional neural network. *Geophys. Res. Lett.* **48**(18), e2021GL094443. https://doi.org/10.1029/2021GL094443 (2021).
26. Noori, R., Karbassi, A. R., Ashrafi, K., Ardestani, M. & Mehrdadi, N. Development and application of reduced-order neural network model based on proper orthogonal decomposition for BOD 5 monitoring: Active and online prediction. *Environ. Prog. Sustain. Energy* **32**(1), 120–127. https://doi.org/10.1002/ep.10611 (2013).
27. Noori, R., Farokhnia, A., Morid, S. & Riahi Madvar, H. Effect of input variables preprocessing in artificial neural network on monthly flow prediction by PCA and wavelet transformation. *J. Water Wastewater* **69**, 13–22 (2009) (**In Persian**).
28. Tayfur, G. & Singh, V. P. Predicting longitudinal dispersion coefficient in natural streams by artificial neural network. *J. Hydraul. Eng.* **131**(11), 991–1000. https://doi.org/10.1061/(ASCE)0733-9429(2005)131:11(991) (2005).
29. Toprak, Z. F., Hamidi, N., Kisi, O. & Gerger, R. Modeling dimensionless longitudinal dispersion coefficient in natural streams using artificial intelligence methods. *KSCE J. Civ. Eng.* **18**(2), 718–730. https://doi.org/10.1007/s12205-014-0089-y (2014).
30. Parsaie, A., Emamgholizadeh, S., Azamathulla, H. M. & Haghiabi, A. H. ANFIS-based PCA to predict the longitudinal dispersion coefficient in rivers. *Int. J. Hydrol. Sci. Technol.* **8**(4), 410–424. https://doi.org/10.1504/IJHST.2018.095537 (2018).
31. Azar, N. A., Milan, S. G. & Kayhomayoon, Z. The prediction of longitudinal dispersion coefficient in natural streams using LS-SVM and ANFIS optimized by Harris hawk optimization algorithm. *J. Contam. Hydrol.* **240**, 103781. https://doi.org/10.1016/j.jconhyd.2021.103781 (2021).
32. Noori, R. *et al.* Assessment of input variables determination on the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction. *J. Hydrol.* **401**(3–4), 177–189. https://doi.org/10.1016/j.jhydrol.2011.02.021 (2011).
33. Tayfur, G. Fuzzy, ANN, and regression models to predict longitudinal dispersion coefficient in natural streams. *Hydrol. Res.* **37**(2), 143–164. https://doi.org/10.2166/nh.2006.0012 (2006).
34. Toprak, Z. F. & Cigizoglu, H. K. Predicting longitudinal dispersion coefficient in natural streams by artificial intelligence methods. *Hydrol. Process. Int. J.* **22**(20), 4106–4129. https://doi.org/10.1002/hyp.7012 (2008).
35. Toprak, Z. F. & Savci, M. E. Longitudinal dispersion coefficient modeling in natural channels using fuzzy logic. *Clean: Soil, Air, Water* **35**(6), 626–637. https://doi.org/10.1002/clen.200700122 (2007).
36. Piotrowski, A. P., Rowinski, P. M. & Napiorkowski, J. J. Comparison of evolutionary computation techniques for noise injected neural network training to estimate longitudinal dispersion coefficients in rivers. *Expert Syst. Appl.* **39**(1), 1354–1361. https://doi.org/10.1016/j.eswa.2011.08.016 (2012).
37. Sahay, R. R. Predicting longitudinal dispersion coefficients in sinuous rivers by genetic algorithm. *J. Hydrol. Hydromech.* **61**(3), 214. https://doi.org/10.2478/johh-2013-0028 (2013).
38. Najafzadeh, M. & Tafarojnoruz, A. Evaluation of neuro-fuzzy GMDH-based particle swarm optimization to predict longitudinal dispersion coefficient in rivers. *Environ. Earth Sci.* **75**(2), 157. https://doi.org/10.1007/s12665-015-4877-6 (2016).
39. Noori, R., Ghiasi, B., Sheikhian, H. & Adamowski, J. F. Estimation of the dispersion coefficient in natural rivers using a granular computing model. *J. Hydraul. Eng.* **143**(5), 04017001. https://doi.org/10.1061/(ASCE)HY.1943-7900.0001276 (2017).
40. Kargar, K. *et al.* Estimating longitudinal dispersion coefficient in natural streams using empirical models and machine learning algorithms. *Eng. Appl. Comput. Fluid Mech.* **14**(1), 311–322. https://doi.org/10.1080/19942060.2020.1712260 (2020).
41. Riahi-Madvar, H., Dehghani, M., Parmar, K. S., Nabipour, N. & Shamshirband, S. Improvements in the explicit estimation of pollutant dispersion coefficient in rivers by subset selection of maximum dissimilarity hybridized with ANFIS-firefly algorithm (FFA). *IEEE Access* **8**, 60314–60337. https://doi.org/10.1109/ACCESS.2020.2979927 (2020).
42. Noori, R., Deng, Z., Kiaghadi, A. & Kachoosangi, F. T. How reliable are ANN, ANFIS, and SVM techniques for predicting longitudinal dispersion coefficient in natural rivers?. *J. Hydraul. Eng.* **142**(1), 04015039. https://doi.org/10.1061/(ASCE)HY.1943-7900.0001062 (2016).
43. Ghiasi, B., Sheikhian, H., Zeynolabedin, A. & Niksokhan, M. H. Granular computing–neural network model for prediction of longitudinal dispersion coefficients in rivers. *Water Sci. Technol.* **80**(10), 1880–1892. https://doi.org/10.2166/wst.2020.006 (2019).
44. Montanari, A. & Brath, A. A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. *Water Resour. Res.* **40**(1), W01106. https://doi.org/10.1029/2003WR002540 (2004).
45. Beven, K. & Binley, A. The future of distributed models: Model calibration and uncertainty prediction. *Hydrol. Process.* **6**(3), 279–298. https://doi.org/10.1002/hyp.3360060305 (1992).
46. Feyen, L., Vrugt, J. A., Nualláin, B. Ó., van der Knijff, J. & De Roo, A. Parameter optimisation and uncertainty assessment for large-scale streamflow simulation with the LISFLOOD model. *J. Hydrol.* **332**(3–4), 276–289. https://doi.org/10.1016/j.jhydrol.2006.07.004 (2007).
47. Renard, B., Kavetski, D., Kuczera, G., Thyer, M. & Franks, S. W. Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resour. Res.* **46**(5), W05521. https://doi.org/10.1029/2009WR008328 (2010).
48. McMillan, H., Jackson, B., Clark, M., Kavetski, D. & Woods, R. Rainfall uncertainty in hydrological modelling: An evaluation of multiplicative error models. *J. Hydrol.* **400**(1–2), 83–94. https://doi.org/10.1016/j.jhydrol.2011.01.026 (2011).
49. Dobler, C., Hagemann, S., Wilby, R. L. & Stötter, J. Quantifying different sources of uncertainty in hydrological projections in an Alpine watershed. *Hydrol. Earth Syst. Sci.* **16**(11), 4343–4360. https://doi.org/10.5194/hess-16-4343-2012 (2012).
50. Hattermann, F. F. *et al.* Sources of uncertainty in hydrological climate impact assessment: A cross-scale study. *Environ. Res. Lett.* **13**(1), 015006. https://doi.org/10.1088/1748-9326/aa9938 (2018).
51. Moges, E., Demissie, Y., Larsen, L. & Yassin, F. Review: Sources of hydrological model uncertainties and advances in their analysis. *Water* **13**(1), 28. https://doi.org/10.3390/w13010028 (2020).
52. Sheikhian, H., Delavar, M. R. & Stein, A. A GIS-based multi-criteria seismic vulnerability assessment using the integration of granular computing rule extraction and artificial neural networks. *Trans. GIS* **21**(6), 1237–1259. https://doi.org/10.1111/tgis.12274 (2017).
53. Yao, Y. A partition model of granular computing. In *Transactions on Rough Sets I* 232–253 (Springer, 2004). https://doi.org/10.1007/978-3-540-27794-1_11.
54. Sheikhian, H., Delavar, M. R. & Stein, A. Integrated estimation of seismic physical vulnerability of Tehran using rule based granular computing. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **40**(3), 187. https://doi.org/10.5194/isprsarchives-XL-3-W3-187-2015 (2015).
55. Khamespanah, F., Delavar, M. R., Moradi, M. & Sheikhian, H. A GIS-based multi-criteria evaluation framework for uncertainty reduction in earthquake disaster management using granular computing. *Geod. Cartogr.* **42**(2), 58–68. https://doi.org/10.3846/20296991.2016.1199139 (2016).
56. Noori, R. *et al.* Granular computing for prediction of scour below spillways. *Water Resour. Manag.* **31**(1), 313–326. https://doi.org/10.1007/s11269-016-1526-0 (2017).
57. Deng, Z. Q., Singh, V. P. & Bengtsson, L. Longitudinal dispersion coefficient in straight rivers. *J. Hydraul. Eng.* **127**(11), 919–927. https://doi.org/10.1061/(ASCE)0733-9429(2001)127:11(919) (2001).

58. Barati Moghaddam, M., Mazaheri, M. & MohammadVali Samani, J. A comprehensive one-dimensional numerical model for solute transport in rivers. *Hydrol. Earth Syst. Sci.* **21**(1), 99–116. https://doi.org/10.5194/hess-21-99-2017 (2017).
59. Kilpatrick, F. A. & Wilson, J. F. *Measurement of Time of Travel in Streams by Dye Tracing* vol. 3. (US Government Printing Office, 1989).
60. Iwasa, Y. & Aya, S. Transverse mixing in a river with complicated channel geometry. *Bull. Disaster Prev. Res. Inst.* **41**(3), 129–175 (1991).
61. Liu, H. Predicting dispersion coefficient of streams. *J. Environ. Eng. Div.* **103**(1), 59–69. https://doi.org/10.1061/JEEGAV.0000605 (1977).
62. Smith, R. 'Physics of Dispersion' coastal and estuarine pollution—methods and solutions' technical sessions, *Scottish Hydraulic Study Group*, One day seminar 3rd Aprill, Glasgow (1992).
63. Taylor, G. I. The dispersion of matter in turbulent flow through a pipe. *Proc. R. Soc. Lond. Ser. A Math. Phys. Sci.* **223**(1155), 446–468. https://doi.org/10.1098/rspa.1954.0130 (1954).
64. Taylor, G. I. Dispersion of soluble matter in solvent flowing slowly through a tube. *Proc. R. Soc. Lond. Ser. A Math. Phys. Sci.* **219**(1137), 186–203. https://doi.org/10.1098/rspa.1953.0139 (1953).
65. Elder, J. The dispersion of marked fluid in turbulent shear flow. *J. Fluid Mech.* **5**(4), 544–560. https://doi.org/10.1017/S0022112059000374 (1959).
66. Fischer, H. B. Longitudinal dispersion in laboratory and natural streams. *Calif. Inst. Technol.* https://doi.org/10.7907/Z9F769HC (1966).
67. Carr, M. L. & Rehmann, C. R. Measuring the dispersion coefficient with acoustic Doppler current profilers. *J. Hydraul. Eng.* **133**(8), 977–982. https://doi.org/10.1061/(ASCE)0733-9429(2007)133:8(977) (2007).
68. Papadimitrakis, I. & Orphanos, I. Longitudinal dispersion characteristics of rivers and natural streams in Greece. *Water Air Soil Pollut. Focus* **4**(4), 289–305. https://doi.org/10.1023/B:WAFO.0000044806.98243.97 (2004).
69. Fischer, H. B., List, J. E., Koh, C. R., Imberger, J. & Brooks, N. H. *Mixing in Inland and Coastal Waters* (Academic Press, 1979).
70. Balf, M. R., Noori, R., Berndtsson, R., Ghaemi, A. & Ghiasi, B. Evolutionary polynomial regression approach to predict longitudinal dispersion coefficient in rivers. *J. Water Supply Res. Technol. AQUA* **67**(5), 447–457. https://doi.org/10.2166/aqua.2018.021 (2018).
71. Riahi-Madvar, H., Dehghani, M., Seifi, A. & Singh, V. P. Pareto optimal multigene genetic programming for prediction of longitudinal dispersion coefficient. *Water Resour. Manag.* **33**(3), 905–921. https://doi.org/10.1007/s11269-018-2139-6 (2019).
72. Calandro, A. J. Time of travel of solutes in Louisiana streams. Louisiana Department of Public Works Water Resources Technical Report (No. 17). Accessed 14 Oct 2020. https://wise.er.usgs.gov/dp/pdfs/TR17.pdf (USGS, 1978).
73. Yao, Y.Y. On modeling data mining with granular computing. In *25th Annual International Computer Software and Applications Conference. COMPSAC 2001* 638–643. (IEEE, 2001). https://doi.org/10.1109/CMPSAC.2001.960680.
74. Yao, J. T. & Yao, Y. Y. Induction of classification rules by granular computing. In *International Conference on Rough Sets and Current Trends in Computing* 331–338. (Springer, 2002). https://doi.org/10.1007/3-540-45813-1_43.
75. Yao, Y. Y. & Zhong, N. Granular computing using information tables. In *Data Mining, Rough Sets and Granular Computing. Studies in Fuzziness and Soft Computing*, vol. 95 (eds. Lin T. Y., Yao Y. Y. & Zadeh L.A.) (Physica, 2002). https://doi.org/10.1007/978-3-7908-1791-1_5.
76. Pawlak, Z. Rough sets. *Int. J. Comput. Inf. Sci.* **11**(5), 341–356. https://doi.org/10.1007/BF01001956 (1982).
77. Haykin, S. *Neural Networks and Learning Machines* 3rd edn. (Prentice Hall, 2008).
78. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (CRC Press, 1994).
79. Srivastav, R. K., Sudheer, K. P. & Chaubey, I. A simplified approach to quantifying predictive and parametric uncertainty in artificial neural network hydrologic models. *Water Resour. Res.* https://doi.org/10.1029/2006WR005352 (2007).
80. Abbaspour, K. C. *et al.* Modelling hydrology and water quality in the pre-alpine/alpine Thur watershed using SWAT. *J. Hydrol.* **333**(2–4), 413–430. https://doi.org/10.1016/j.jhydrol.2006.09.014 (2007).
81. Bello, R. *et al.* (eds) *Granular Computing: At the Junction of Rough Sets and Fuzzy Sets* Vol. 224 (Springer, 2007).
82. Koh, Y. S. & Rountree, N. (eds) *Rare Association Rule Mining and Knowledge Discovery: Technologies for Infrequent and Critical Event Detection: Technologies for Infrequent and Critical Event Detection* Vol. 3 (IGI Global, 2009).
83. Etemad-Shahidi, A. & Taghipour, M. Predicting longitudinal dispersion coefficient in natural streams using M5′ model tree. *J. Hydraul. Eng.* **138**(6), 542–554. https://doi.org/10.1061/(ASCE)HY.1943-7900.0000550 (2012).
84. Sahay, R. R. & Dutta, S. Prediction of longitudinal dispersion coefficients in natural rivers using genetic algorithm. *Hydrol. Res.* **40**(6), 544–552. https://doi.org/10.2166/nh.2009.014 (2009).
85. Najafzadeh, M. *et al.* A comprehensive uncertainty analysis of model-estimated longitudinal and lateral dispersion coefficients in open channels. *J. Hydrol.* https://doi.org/10.1016/j.jhydrol.2021.126850 (2021).
86. Dehghani, M., Zargar, M., Riahi-Madvar, H. & Memarzadeh, R. A novel approach for longitudinal dispersion coefficient estimation via tri-variate archimedean copulas. *J. Hydrol.* **584**, 124662. https://doi.org/10.1016/j.jhydrol.2020.124662 (2020).
87. Memarzadeh, R. *et al.* A novel equation for longitudinal dispersion coefficient prediction based on the hybrid of SSMD and whale optimization algorithm. *Sci. Total Environ.* **716**, 137007. https://doi.org/10.1016/j.scitotenv.2020.137007 (2020).
88. Noori, R., Karbassi, A., Farokhnia, A. & Dehghani, M. Predicting the longitudinal dispersion coefficient using support vector machine and adaptive neuro-fuzzy inference system techniques. *Environ. Eng. Sci.* **26**(10), 1503–1510. https://doi.org/10.1089/ees.2008.0360 (2009).

Author contributions

Data collection and analysis were carried out by B.G., R.N, H.S. and A.Z. R.N. conceived the study conceptually. The models were ran by R.N. and H.S. The first draft of manuscript was prepared by B.G., R.N., Y.S., C.J. and M.H. The funding acquisition was made by C.J. and M.H. The analyses and results were supervised and validated by R.N., S.M.B. and S.A. All authors read and approved the final version of manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to R.N. or M.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.