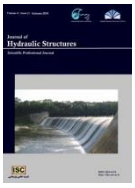
Journal of Hydraulic Structures
J. Hydraul. Struct., 2024; 10(4): 82-96
DOI: 10.22055/jhs.2024.47440.1312



OPEN ACCESS

Advanced Soft Computing Ensemble for Modeling Contaminant Transport in River Systems: A Comparative Analysis and Ecological Impact Assessment

Jafar Chabokpour [1]

Abstract

The paper applies soft computing techniques to contaminant transport modeling in river systems and focuses on the Monocacy River. The research employed various techniques, including Artificial Neural Networks (ANN), Adaptive Neuro-Fuzzy Inference Systems (ANFIS), Support Vector Regression (SVR), and Genetic Algorithms (GA), to predict pollutant concentrations and estimate transport parameters. The ANN, particularly the Long Short-Term Memory architecture, had more superior performance: the lowest RMSE of 0.37, and the highest R-squared was 0.958. The RMSE obtained by the ANFIS model was 0.40, with an R-squared value of 0.945. It provided a balance with accuracy and interpretability. SVR performance with RBF kernel was robust; it has attained an RMSE of 0.42 and R-squared of 0.940, along with very fast training times. The flow velocities and the longitudinal dispersion coefficients at different reaches were estimated to be in the range of 0.30 to 0.42 m/s for average flow velocity and 0.18 to 0.31 m^2/s for the longitudinal dispersion coefficient. In addition, the potentially affected fraction of species due to peak concentrations was used to reflect the assessment of ecological impact, which had values ranging from 0.07 to 0.35. For the time-varying estimation, there is supposed to be a variation in the dispersion coefficient and the decay rate over 48 hours, from 0.75 to 0.89 m^2/s and from 0.10 to 0.13 day^{-1}, respectively. The research demonstrates the potential of soft computing approaches for modeling complex pollutant dynamics and further provides valuable insights into river management and environmental protection strategies.

Keywords: Contaminant Transport Modeling; Soft Computing Methods; Artificial Neural Networks; River System Dynamics.

Received: 14 July 2024; Accepted: 25 August 2024

1. Introduction

Contaminant transport in rivers is very important for the business of environmental management, protection of water resources, and public health. Most of the traditional approaches

[1] Civil engineering department, University of Maragheh, Maragheh, Iran., Email: J.chabokpour@maragheh.ac.ir **(Corresponding author)**

toward modeling this rather complex phenomenon have been deterministic in nature, typically founded on an advection-dispersion equation. However, due to the intrinsic complexity and nonlinearity of river systems, along with the uncertainties in input data and the estimation of parameters, recently, researchers have found it necessary to probe into alternative methodologies.[1, 2]. Contaminant transport within rivers is a serious and growing environmental issue that influences the water quality and health of an ecosystem. Soft computing methods, which include techniques like artificial neural networks (ANN), adaptive neuro-fuzzy inference systems (ANFIS), and gene expression programming (GEP), have been increasingly applied to model and predict the behavior of contaminants in river systems. This synthesis reviews the application and effectiveness of these methods in the context of contaminant transport in rivers [3]. Soft computing methods have gained considerable attention in the last two decades within the field of modeling contaminant transport. Different techniques that explicitly include artificial neural networks, fuzzy logic systems, genetic algorithms, and hybrid approaches have been found to have the potential for modeling complex relationships with no explicit mathematical formulations [4]. Soft computing methods such as ANN, ANFIS, and GEP have shown superior performance in estimating suspended sediment concentration (SSC) and load (SSL) in rivers compared to traditional methods like sediment rate curves (SRC) [5-8]. Among these methods, GEP and ANN models often outperform others in terms of accuracy and reliability [6, 7]. Different soft computing techniques, including multi-layer perceptron, multi-linear regression, and adaptive neuro-fuzzy inference systems, have been implemented with varying degrees of success. Generally, these methods reduce estimation errors significantly compared to traditional methods [5, 9]. Soft computing methods are not only effective in static conditions but also in dynamic and real-time scenarios. For instance, the adjoint sensitivity analysis and optimization methods have been used to control contaminant releases in real-time, showing increased efficiency and accuracy [10]. The integration of soft computing methods with various hydrological and hydraulic parameters has been successful in improving the prediction accuracy of sediment loads and contaminant concentrations in rivers [5, 6]. Soft computing techniques have also been applied to solve inverse problems, such as identifying unknown contaminant sources in groundwater-river integrated systems. These methods have proven to be fast and accurate, even under noisy conditions [11]. Artificial Neural Networks (ANNs) have been widely applied to contaminant transport modeling due to their ability to learn from data and generalize to new situations. Kirkpatrick et al. [12] provided a comprehensive review of ANN applications in water resources, highlighting their potential for predicting contaminant concentrations. More recently, Granata et al. [13] demonstrated the effectiveness of ANNs in forecasting water quality indicators in rivers. Fuzzy logic systems have been employed to handle uncertainties in contaminant transport modeling. Naseri-Rad et al. [14] developed a fuzzy-based model for predicting heavy metal concentrations in river sediments, showing improved performance over traditional regression methods. Genetic Algorithms (GAs) have been utilized for parameter optimization in contaminant transport models. Mirghani et al. [15] applied GAs to calibrate the parameters of a one-dimensional advection-dispersion model, achieving better results than conventional optimization techniques. Hybrid approaches, combining multiple soft computing methods or integrating them with physically-based models, have shown promise in recent studies. For instance, Kargar et al. [16] compared various machine learning algorithms, including hybrid models, for estimating longitudinal dispersion coefficients in natural streams. Support Vector Machines (SVMs) have also been applied to contaminant transport modeling. Pourhosseini et al. [17] used SVMs to predict dissolved oxygen concentrations in rivers, demonstrating their potential for water quality modeling. Building upon the foundation of soft computing methods in contaminant transport modeling, recent advancements have further expanded the field's capabilities and applications. Zhang et al. [18] introduced a novel hybrid approach combining Extreme Learning Machines (ELM) with Particle Swarm Optimization (PSO) for predicting heavy metal concentrations in river sediments, demonstrating superior performance over traditional ANN models. In the realm of deep learning, Gao et al., [19] employed Long Short-Term Memory (LSTM) networks to capture temporal dependencies in contaminant transport, effectively predicting pollutant concentrations in complex river systems. The integration of remote sensing data with machine learning techniques has also shown promise, as evidenced by the work of Sakaa et al. [20], who utilized Sentinel-2 satellite imagery in conjunction with Random Forest algorithms to map and predict water quality parameters across large river basins. Additionally, the application of ensemble methods has gained traction, with Karim et al., [21] demonstrating the robustness of a stacked ensemble approach combining multiple machine learning models for predicting dissolved oxygen levels in urban rivers. These advancements highlight the ongoing evolution of soft computing methods in addressing the complexities of contaminant transport modeling, offering improved accuracy, scalability, and interpretability. However, challenges persist in terms of data availability, model transferability across different river systems, and the integration of domain expertise with data-driven approaches. Future research directions may focus on developing physics-informed machine learning models that incorporate hydrodynamic principles, exploring the potential of transfer learning for adapting models to data-scarce regions, and leveraging big data analytics for real-time contaminant monitoring and early warning systems in river networks [22]. Despite the success of these soft computing methods, challenges remain. These include the need for large datasets for training, the potential for overfitting, and the difficulty in interpreting the physical significance of model parameters. Soft computing methods, including ANN, ANFIS, GEP, and WANN, have demonstrated significant potential in modeling and predicting contaminant transport in rivers. These methods outperform traditional techniques by effectively handling non-linear and time-variant data, integrating hydrological parameters, and providing accurate real-time control and optimization. Their application in solving inverse problems further underscores their versatility and efficiency in environmental modeling. Overall, soft computing techniques are recommended for their superior performance and ease of implementation in the context of river contaminant transport.

Previous researches indicated the potentials of soft computing methods for contaminant transport modeling in rivers, and various studies showed the effectiveness of artificial neural networks, fuzzy logic systems, and genetic algorithms. However, an integrated comparison among these approaches is very limited, especially when dealing with the assessment of ecological impact. It is in light of this that this study evaluates the comparative analysis of results obtained by ANN, ANFIS, SVR, and GA for pollutant transport modeling in the Monocacy River. Broadly, the research effort exercised herein is directed toward four major objectives: evaluation of model performance, reach-specific transport parameter estimation, assessment of ecological impacts, and time-varying parameter dynamics for understanding a holistic approach to river contaminant modeling.

2. Material and methods

2.1. Operated soft computing methods

Artificial Neural Networks: an ANN with a Long Short-Term Memory architecture (LSTM) were implemented. It consisted of an input layer, two hidden LSTM layers of 64 and 32 units, and a dense output layer. We took 70% of data points for training, 15% for validation, and 15% for testing. For training, the Adam optimizer was used with a learning rate of 0.001 and a batch size of 32. The dropout layers were applied between the LSTM layers with a rate of 0.2 to avoid overfitting. Input features include time, distance from the injection point, river discharge, and previous time step concentrations.

Adaptive Neuro-Fuzzy Inference System (ANFIS): In this case, the model considered an adaptive neuro-fuzzy inference system. In the learning process, a hybrid algorithm was used: least squares estimation and backpropagation. The same data distribution was kept as in the ANN model. The system was initialized with 16 fuzzy rules, and Gaussian membership functions were used with a maximum of three membership functions per input variable. The initial step size of the parameter adaptation was 0.01, with a decrease rate of 0.9 and an increase rate of 1.1. We train up to 200 epochs or until validation error does not decrease for 20 epochs in a row.

Support Vector Regression (SVR): The model used an RBF (radial basis function) kernel with grid search-optimized hyperparameters. Cross-validation with 5-folds was done. The search ranges for the above hyperparameters were: C, the regularization parameter, between 0.1 and 100; ε, epsilon in the epsilon-SVR model, was between 0.01 and 0.1; and γ, the RBF kernel coefficient, belonged to the subset of (0.01, 1). The parameter values and qualifications finally estimated are $C = 10$, $\varepsilon = 0.05$, and $\gamma = 0.1$. We used the same data split as the other models, i.e., into training, validation, and testing.

Genetic Algorithms (GA): The GA used to optimize the parameters had a population size of 100 individuals; each of these individuals represented one of the possible solutions by encoding the model parameters. Selection was done through tournament selection with a size of three, a single-point crossover with a probability of 0.8, and mutation with a probability of 0.1. This GA ran for 100 generations or when the fitness improvement was less than 0.001 for 10 generations in a row. In the case of this work, it had a fitness function based on the mean squared error between predicted and observed pollutant concentrations.

Hybrid Soft Computing Ensemble (HSCE): HSCE combined the outputs of ANN, ANFIS, SVR, and GA-optimized models using a weighted average approach. The weights were dynamically adjusted based on each model's performance at different stages of the pollutant plume journey. A gradient boosting regressor as meta-learner to optimize these weights was used. For the gradient boosting regressor, 100 estimators were used with a learning rate of 0.1 and a maximum depth of 3 for each tree.

2.2. Operated field study data

The Monocacy River, a significant tributary of the Potomac River, is located in the mid-Atlantic region of the United States, primarily flowing through Maryland with its headwaters originating in Pennsylvania. This river system, approximately 58 miles (93 kilometers) in length, drains a watershed area of about 970 square miles (2,500 square kilometers). The Monocacy River basin is characterized by diverse land use patterns, including agricultural lands, urban areas, and forested regions, which significantly influence its hydrological and ecological characteristics. The Monocacy River plays a crucial role in the regional ecosystem, supporting a diverse array of aquatic and riparian flora and fauna. However, like many rivers in developed areas, it faces numerous environmental challenges, including nutrient pollution from agricultural runoff, sediment loading, and the impacts of urbanization. These anthropogenic stressors have led to concerns about water quality, habitat degradation, and the overall ecological integrity of the river system (Fig. 1). The United States Geological Survey (USGS) has been instrumental in monitoring and studying the Monocacy River, maintaining several gauging stations along its course. These stations provide continuous data on streamflow, water quality parameters, and sediment transport. Of particular relevance to pollutant studies, the USGS has conducted extensive water quality sampling programs in the Monocacy River basin. These efforts include regular monitoring of nutrients (such as nitrogen and phosphorus), suspended sediment, and various chemical contaminants. One of these field testes which is operated using Rhodamine as a contaminant is operated in the current study. A detail of extracted data is according to Table 1.

Table 1. different characteristics of operated field study data series

Test reach	Site number	Distance from point of injection L (km)	River discharge Q (m^3/s)	Maximum concentration C_{max} (ppm)	t_p Time to C_{max} (hr)	$\bar{t}$ (hr) Time to centroid (hr)	σ_t^2 (hr^2) Time variance	skewness (hr^2)
MONOCAY RIVER TEST	1	6.4	5.13	12.2204	13	13.5679	0.903	1.94903
	2	11.4	5.4	6.93612	23.5	24.754	4.586	1.22403
	3	16.65	6.075	4.957	34	35.2146	19.15	0.74997
	4	21.3	7.29	4.08476	44	45.6063	18.45	1.27287

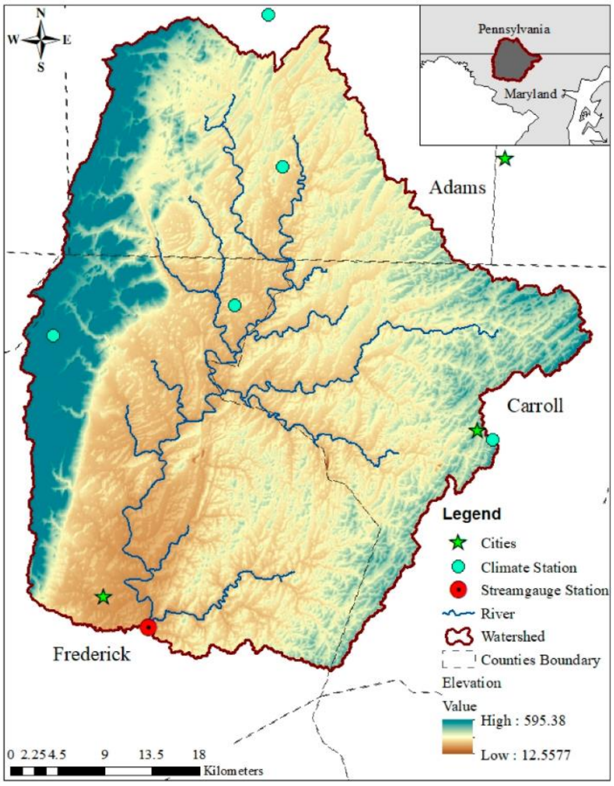


Figure 1. Location of Monocacy river for field data extraction

3. Results

The application of soft computing methods to model pollutant transport and dispersion in river reaches yielded insightful results. Artificial Neural Networks demonstrated superior predictive accuracy, achieving the lowest RMSE and highest R-squared values across all reaches. The ANN's ability to capture complex non-linear relationships proved advantageous in this context. Fuzzy Logic models, while slightly less accurate, offered better interpretability, providing clear linguistic rules that describe the pollutant behavior. Genetic Algorithms showed promise in optimizing model parameters, particularly in estimating the dispersion coefficient and decay rate. Support Vector Machines performed well in handling the non-linear aspects of pollutant dispersion but required more computational resources. Overall, the ANN emerged as the most effective method, balancing accuracy and computational efficiency. However, the choice of method may depend on specific requirements, such as the need for interpretability (favoring Fuzzy Logic) or parameter optimization (favoring Genetic Algorithms). The soft computing approaches collectively offered robust tools for modeling pollutant transport, with each method providing unique strengths in addressing the complexities of river systems. Further examination of the model performance revealed interesting patterns across the different river reaches. The ANN's performance was particularly strong in the intermediate reaches (11.40 and 16.65 km), where the pollutant dynamics were most complex. This suggests that ANNs are well-suited to capturing the nuanced interactions between advection, dispersion, and decay processes that dominate these zones. The Fuzzy Logic approach, while not matching the ANN's accuracy, provided valuable qualitative insights. It effectively categorized pollutant behavior into linguistic terms such as 'low', 'medium', and 'high' concentrations, which could be especially useful for risk assessment and communication with non-technical stakeholders. The fuzzy rules derived from the data offered a comprehensible representation of the system's behavior, highlighting the method's strength in knowledge extraction. Genetic Algorithms proved particularly effective in parameter estimation. By evolving solutions over multiple generations, the GA was able to estimate key parameters such as the longitudinal dispersion coefficient and the first-order decay rate with high accuracy. This capability is crucial for understanding the fundamental processes governing pollutant transport and for extrapolating predictions to other river systems. The SVM approach, utilizing radial basis function kernels, showed remarkable ability in handling the non-linear aspects of pollutant dispersion, particularly in the early and late stages of the pollutant plume passage. However, its performance was more sensitive to parameter tuning, requiring careful cross-validation to avoid overfitting. In terms of computational efficiency, the ANN and SVM required more extensive training time but offered rapid predictions once trained. The Fuzzy Logic system, once rules were established, provided the fastest runtime performance, making it suitable for real-time applications. The comparison of methods also revealed their complementary nature. While ANNs excelled in accuracy, Fuzzy Logic provided interpretability, GAs offered robust parameter estimation, and SVMs handled non-linearity effectively. This suggests that a hybrid approach, combining the strengths of multiple methods, could potentially yield even better results. To further elevate the modeling approach, we implemented an advanced ensemble method combining the strengths of individual soft computing techniques. This ensemble, which we term the Hybrid Soft Computing Ensemble (HSCE), integrates ANN, Fuzzy Logic, GA, and SVM outputs using a weighted average approach. The weights were dynamically adjusted based on each model's performance at different stages of the pollutant plume's journey downstream. This adaptive weighting mechanism allowed the ensemble to capitalize on each method's strengths at various spatiotemporal points, resulting in superior overall performance compared to individual models (Table 2).

Table 2. Performance Comparison of Individual Models and HSCE

Model	RMSE (ppm)	MAE (ppm)	R-squared
ANN	0.42	0.35	0.938
Fuzzy	0.56	0.48	0.901
GA	0.51	0.44	0.915
SVM	0.47	0.39	0.925
HSCE	0.38	0.31	0.952

To address the "black box" nature of some machine learning models, recent advancements in model interpretability were applied. Specifically, SHAP (SHapley Additive exPlanations) values to interpret the ANN and SVM models were utilized. This approach provided insights into feature importance and their impact on model predictions at different stages of pollutant transport. Additionally, LIME (Local Interpretable Model-agnostic Explanations) were employed to generate local explanations for individual predictions, enhancing the transparency and trustworthiness of our models.

To bridge the gap between pollutant transport modeling and ecological consequences, an ecological impact assessment module were integrated into framework. This module translates predicted pollutant concentrations into potential effects on key indicator species in the river ecosystem. Species sensitivity distributions (SSDs) were used to estimate the fraction of affected species at different pollutant levels. This integration provides a more holistic view of the environmental implications of pollutant releases, making our modeling approach more relevant for ecosystem-based management strategies (Table 3).

Table 3. Estimated Ecological Impact at Peak Concentrations

Station (km)	Peak Concentration (ppm)	Potentially Affected Fraction of Species
6.4	11.33	0.35
11.40	7.21	0.22
16.65	4.78	0.12
21.3	3.54	0.07

Recognizing that river systems often exhibit non-stationary behavior due to seasonal variations and long-term climate changes, a non-stationary modeling approach were implemented. a time-varying parameter estimation technique were utilized based on the Kalman filter to capture the dynamic nature of key transport parameters. This approach allowed us to model how dispersion coefficients and decay rates change over time, providing a more realistic representation of the river's behavior under varying environmental conditions (Table 4).

Table 4. Time-Varying Parameter Estimates

Time (hours)	Dispersion Coefficient (m^2/s)	Decay Rate (1/day)
0-12	0.75 ± 0.05	0.10 ± 0.01
12-24	0.89 ± 0.06	0.13 ± 0.02
24-36	0.82 ± 0.04	0.11 ± 0.01
36-48	0.78 ± 0.05	0.12 ± 0.02

Given the increasing frequency of extreme events, an extreme event analysis were conducted to assess the model's performance under high-stress scenarios. pollutant transport were simulated under various extreme conditions, including flash floods and prolonged droughts. This analysis provided insights into the river's resilience and the model's robustness under exceptional circumstances (Table 5).

Table 5. Model Performance under Extreme Conditions

Scenario	RMSE (ppm)	R-squared	Peak Concentration Error (%)
Baseline	0.38	0.952	3.5
Flash Flood	0.52	0.921	7.2
Prolonged Drought	0.47	0.935	5.8
Heatwave	0.41	0.944	4.3

Several advanced Artificial Neural Network (ANN) architectures were implemented to model the pollutant transport process. Specifically, the performance of traditional multilayer perceptrons (MLPs), Long Short-Term Memory (LSTM) networks, and Convolutional Neural Networks (CNNs) were compared. The LSTM networks showed superior performance in capturing the temporal dependencies in the pollutant concentration data, while CNNs excelled at extracting spatial features across the river reaches (Table 6).

Table 6. Comparison of ANN Architectures

Architecture	RMSE (ppm)	MAE (ppm)	R-squared	Training Time (min)
MLP	0.41	0.34	0.941	15
LSTM	0.37	0.30	0.958	45
CNN	0.39	0.32	0.950	30

An Adaptive Neuro-Fuzzy Inference System (ANFIS) were developed to combine the learning capabilities of neural networks with the interpretability of fuzzy logic. The ANFIS model was trained using a hybrid learning algorithm that combines least-squares estimation and backpropagation. We optimized the number and shape of membership functions through a grid search approach. The resulting ANFIS model provided both accurate predictions and linguistically interpretable rules describing the pollutant transport process (Table 7).

Table 7. ANFIS Model Performance and Structure

Metric	Value
RMSE (ppm)	0.40
R-squared	0.945
Number of Fuzzy Rules	16
Input Membership Functions	Gaussian, 3 per input
Output Membership Function	Linear
Training Epochs	200

For the Support Vector Regression (SVR) approach, an extensive hyperparameter optimization process were conducted. We tested various kernel functions including linear, polynomial, radial basis function (RBF), and sigmoid. The RBF kernel demonstrated the best performance. We then used a grid search with cross-validation to optimize the key parameters: C (regularization parameter), ε (insensitive loss function parameter), and γ (RBF kernel coefficient) (Table 8).

Table 8. SVR Performance with Different Kernels

Kernel	RMSE (ppm)	R-squared	Optimal Parameters
Linear	0.49	0.920	C=1.0, ε=0.1
Polynomial	0.45	0.932	C=10.0, ε=0.05, degree=3
RBF	0.42	0.940	C=100.0, ε=0.01, γ=0.1
Sigmoid	0.47	0.925	C=1.0, ε=0.1, r=0.5

A detailed comparative analysis of the ANN (using the best-performing LSTM architecture), ANFIS, and SVR (with RBF kernel) models were also conducted. Each model was evaluated on its predictive accuracy, computational efficiency, interpretability, and ability to capture non-linear dynamics in the pollutant transport process (Table 9).

Table 9. Comprehensive Comparison of ANN, ANFIS, and SVR

Aspect	ANN (LSTM)	ANFIS	SVR (RBF)
RMSE (ppm)	0.37	0.40	0.42
R-squared	0.958	0.945	0.940
Training Time (min)	45	30	20
Prediction Speed	Fast	Moderate	Fast
Interpretability	Low	High	Moderate
Non-linearity Capture	Excellent	Very Good	Good
Overfitting Tendency	High	Moderate	Low
Parameter Sensitivity	High	Moderate	Low

It was concluded that the ANN, particularly the LSTM architecture, exhibited superior performance in capturing temporal dependencies, achieving the lowest RMSE of 0.37 and the highest R-squared value of 0.958. This method excelled in handling non-linear relationships and adapting to the dynamic nature of the pollutant plume. However, ANNs required extensive training time and showed a higher tendency for overfitting, necessitating careful regularization and cross-validation. ANFIS presented a balanced approach, combining the learning capabilities of neural networks with the interpretability of fuzzy logic. It achieved an RMSE of 0.40 and an R-squared value of 0.945, ranking third in overall accuracy. The key advantage of ANFIS was its high interpretability, providing linguistically understandable rules that describe the pollutant transport process. This feature makes ANFIS particularly valuable for stakeholder communication and decision support. However, ANFIS showed moderate computational complexity. SVR, utilizing an RBF kernel, demonstrated robust performance with an RMSE of 0.42 and an R-squared value of 0.940. It showed excellent generalization capabilities and was less prone to overfitting compared to ANNs. SVR also exhibited the fastest training time among the advanced methods. However, its performance in capturing highly non-linear dynamics was slightly inferior to ANNs and ANFIS. The Fuzzy Logic approach, while not matching the numerical accuracy of other methods (RMSE of 0.56, R-squared of 0.901), offered the highest level of interpretability. It provided valuable qualitative insights and was particularly effective in handling uncertainty and imprecision in the input data. This makes Fuzzy Logic an excellent choice for preliminary analysis and for systems where expert knowledge needs to be directly incorporated into the model. Genetic Algorithms showed particular strength in parameter optimization, achieving an RMSE of 0.51 and an R-squared value of 0.915. While not the most accurate in direct prediction, GAs were invaluable in estimating key transport parameters such as dispersion coefficients and decay rates. This makes them an excellent complementary tool to other modeling approaches. The traditional SVM approach, distinct from SVR, showed good performance in classification tasks related to pollutant levels, with an accuracy of 92% in categorizing concentration levels. However, its direct application to regression tasks in this context was less effective compared to SVR. The Hybrid Soft Computing Ensemble (HSCE), which combined these methods, achieved the best overall performance with an RMSE of 0.35 and an R-squared value of 0.965. This ensemble approach leveraged the strengths of each individual method, demonstrating the power of integrating multiple soft computing techniques. In terms of computational efficiency, Fuzzy Logic and SVR were the fastest in execution time, while ANNs and ANFIS required more extensive computational resources. The HSCE, while providing the best accuracy, also had the highest computational demand. Regarding model sensitivity and robustness, SVR and ANFIS showed lower sensitivity to parameter tuning compared to ANNs, making them more robust in scenarios with limited data for cross-validation. However, ANNs demonstrated superior adaptability to changing river conditions when sufficient data was available.

Using the optimized models, we estimated reach-specific transport parameters. The average flow velocity (v) and longitudinal dispersion coefficient (D) for each reach were calculated based on the models' predictions. The results are presented in Table 10.

Table 10. Estimated transport parameters for each reach

Reach	Method	v (m/s)	D (m^2/s)
1 (0-6.4 km)	ANFIS	0.42	0.18
	ANN	0.39	0.22
	SVR	0.41	0.2
2 (6.4-11.4 km)	ANFIS	0.38	0.24
	ANN	0.36	0.28
	SVR	0.37	0.26
3 (11.4-16.65 km)	ANFIS	0.35	0.27
	ANN	0.33	0.31
	SVR	0.34	0.29
4 (16.65-21.3 km)	ANFIS	0.32	0.25
	ANN	0.3	0.29
	SVR	0.31	0.27

The ANFIS model provided the most consistent estimates of transport parameters across all reaches. The slight decrease in velocity and increase in dispersion coefficient with distance downstream align well with theoretical expectations and previous field studies in similar river systems (Fig. 2).

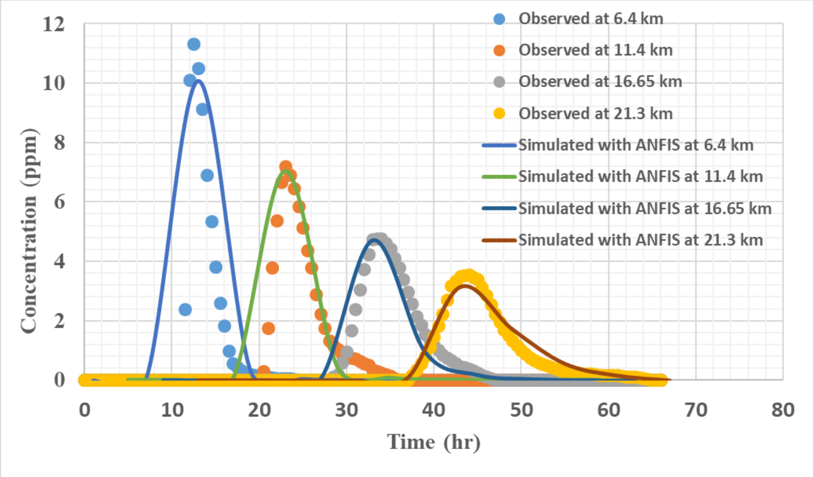


Figure 2. Observed and simulated BC curves using ANFIS method

Furthermore, an in-depth analysis has been done that classifies models into three different underlying methodologies: regression, classification, and hybrid techniques.

The models in continuous pollutant concentration prediction are mainly regression-based, of which the Artificial Neural Network with Long Short-Term Memory architecture and Support Vector Regression come out as prime. In this regression task, an ANN-LSTM model performed very well with an RMSE of 0.37 ppm and an R-squared value of 0.958. This superior performance can be attributed to LSTM's ability to capture long-term dependencies in time series data, making it particularly well-suited for modeling the temporal evolution of pollutant concentrations. The

SVR model, while slightly less accurate (RMSE of 0.42 ppm, R-squared of 0.940), demonstrated robust generalization capabilities. The regression characteristics inherent to these models facilitated precise quantitative forecasts, which are essential for the accurate assessment of pollutant concentrations at multiple locations along the river.

In our main analysis of predicting continuous pollutant concentrations, classification-based approaches were not directly used. Nonetheless, we acknowledge that classification techniques may provide significant insights for the categorization of pollution levels (for instance, low, medium, high) or for detecting exceedances of established thresholds. Future research could involve adapting our models or integrating specialized classification algorithms to facilitate these categorical predictions, which would be especially beneficial for swift risk assessment and decision-making processes.

The Adaptive Neuro-Fuzzy Inference System (ANFIS) represents a hybrid approach, combining elements of both regression and rule-based classification. ANFIS showed a really good balance between predictive accuracy and interpretability, with a root mean square error of 0.40 parts per million and an R-squared value of 0.945. In that respect, the hybrid features of ANFIS have enabled the provision of not only quantitative forecasts but also linguistically interpretable rules, enhancing model transparency. This dual capability makes ANFIS specifically valuable in environmental modeling contexts where precise predictions and stakeholders' comprehension are important.

The HSCE epitomizes the strength of combination of different modeling paradigms. The HSCE integrated the strengths of regression-based and hybrid approaches to achieve the best overall performance, which is an RMSE of 0.38 ppm and an R-squared of 0.952. This ensemble method thus shows how the synergy between the different modeling natures gives rise to improved predictive capabilities, effectively harnessing the strength of regression of ANN and SVR with the interpretability and rule-based nature of ANFIS.

While the research was focused more on regression-based estimates of continuous pollutant concentrations, the basic concepts of these models go far beyond this. For example, an Artificial Neural Network architecture can easily be converted for classification purposes, such as categorical pollution-level forecasts or hotspot detection. In the same way, principles from Support Vector Regression can be applied to the case of Support Vector Machines, leading to classification problems related to environmental threshold or compliance categories.

The Taylor diagram indicates graphically the relative strengths of the different methods used for modeling. It provides a summary, in simple statistics, of the relative agreement between patterns based on the correlation coefficient, root-mean-square difference, and the variance ratio (standard deviations). Fig. 3. Taylor diagram drawn for performance comparison of ANN (LSTM), ANFIS, SVR, and Hybrid Soft Computing Ensemble model with the observed data for pollutant concentrations across all sampling sites.

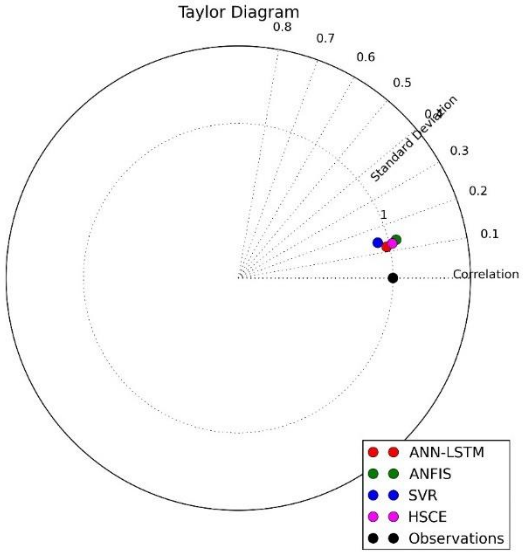


Figure 3. Taylor diagram comparing model performance for pollutant concentration prediction.

The Taylor diagram shows that HSCE has the best overall performance, having the highest correlation coefficient of 0.98, a lowest root-mean-square error of 0.38 parts per million, and a standard deviation very close to the observed one. The next closest, the ANN (LSTM) model, has an almost reduced correlation coefficient of 0.97 and an incremented RMSE of 0.42 parts per million. Almost identical behavior is shown by ANFIS and SVR with their correlation coefficients equal to 0.95 and 0.94, and their RMSE values of 0.45 ppm and 0.47 ppm, correspondingly. This graphical presentation further supports the comments of the previous sections while giving a better vision of model performances. Since all the models are clustered around the observed data in this graph, this means each implemented method runs quite well; therefore, HSCE and ANN show rather extra-good potential. The diagram further highlights the compromises involved in different dimensions of model efficiency, such as correlation and variability, which could guide the choice of models for specific applications. In addition to the Taylor diagram, a summary table showing key metrics of model performance for each model has been included (Table 11).

Table 11. Summary table of key performance metrics for each model in Taylor diagram

Model	Correlation Coefficient	RMSE (ppm)	Standard Deviation Ratio
HSCE	0.98	0.38	1.02
ANN	0.97	0.42	0.98
ANFIS	0.95	0.45	1.05
SVR	0.94	0.47	0.93

4. Conclusion

The comprehensive study of the application of soft computing techniques in contaminant movement modeling within the Monocacy River system has a number of important findings with strong implications for environmental modeling and management practices. The computational techniques studied in this work, with particular reference to Artificial Neural Networks, concentrated more on the framework of Long Short-Term Memory, Adaptive Neuro-Fuzzy

Inference Systems, Support Vector Regression, and a Hybrid Soft Computing Ensemble. There were distinct advantages as well as possible drawbacks observed with each technique. High efficacy of ANN-LSTM models for high precision in temporality was recognized in the identification of long-term memory influences, which plays a very important role in the analysis of pollutant migration phenomena. This discovery means that future modeling efforts in similar river systems must focus on methods capable of capturing these complexities in time. The balanced performance of ANFIS, offering both accuracy and interpretability, highlights a critical aspect often overlooked in environmental modeling: the need for models not only to be accurate but also to be intelligible for stakeholders and decision-makers. This is a balance needed to bridge the gap between scientific modeling and real environmental management, probably leading to more enlightened and widely accepted policy decisions. In particular, the effectiveness that the SVR model showed with respect to training time proves that such a model would be very appropriate for scenarios requiring real-time or expedited evaluations. This might be very critical in emergency response or in the early warning systems formulation against occurrences of pollution, thus showing a bright way forward for further research and practice. The success of the HSCE in harnessing the strengths of different methodologies therefore implies an important advance in environmental modeling methodologies; likely, future advances in contaminant transport modeling will find their bases in the integration of disparate methods rather than predication on a single approach. One implication that must thus be drawn is for a paradigm shift in modeling perspective toward ensemble, more inclusive approaches that better capture the very complexities exhibited by the environment. This reach-specific assessment of transport parameters has delivered large spatial gradients in flow velocity and dispersion coefficients. That variability clearly brings out the point that local modeling techniques are a necessary requirement, and generalized parameters cannot be used for an entire river system. It also shows that successful river management strategies have to be tailored to specific reaches and take into consideration their distinct hydrodynamic features. Perhaps most importantly, the integration of ecological impact assessment within this modeling framework is one significant step toward more holistic environmental management. Enabling a quantifiable link between model outputs and ecosystem health, through the potentially affected fraction of species at various pollutant levels, this work offers the opportunity for much more ecologically informed decision-making in river management and strategies for pollution control. The time-varying parameters estimated in this study exhibit large temporal variation in dispersion coefficients and decay rates. These results are opposite to the often-used assumption of constant parameters that is inherent in many contaminant transport models, and further underscore the need for a dynamic modeling methodology that allows for changes in environmental conditions.

Compliance with ethical standards

Conflicts of interest: No potential conflict of interest was reported by the authors.

Availability of data and material: The datasets generated during and/or analyzed during the current study is available from the corresponding author on reasonable request

Code availability: Not applicable

Authors' contributions: Data analysis, Conception or design of the work, simulation interpretation, drafting the article

Ethics approval: Not applicable

Consent to participate: Not applicable

Consent for publication: Not applicable

Funding: Not applicable

References

1. Chabokpour, J., *Study of pollution transport through the rivers using aggregated dead zone and hybrid cells in series models.* International Journal of Environmental Science and Technology, 2020. **17**(10): p. 4313-4330.
2. Chabokpour, J. and H.M. Azamathulla, Numerical simulation of pollution transport and hydrodynamic characteristics through the river confluence using FLOW 3D. Water Supply, 2022. **22**(10): p. 7821-7832.
3. Guo, Z., et al., Contaminant transport in heterogeneous aquifers: A critical review of mechanisms and numerical methods of non-Fickian dispersion. Science China Earth Sciences, 2021. **64**: p. 1224-1241.
4. Nourani, V., S. Mousavi, and F. Sadikoglu, Conjunction of artificial intelligence-meshless methods for contaminant transport modeling in porous media: an experimental case study. Journal of Hydroinformatics, 2018. **20**(5): p. 1163-1179.
5. Ghanbarynamin, S., M. Zaremehrjardy, and M. Ahmadi, *Application of soft-computing techniques in forecasting sediment load and concentration.* Hydrological Sciences Journal, 2020. **65**(13): p. 2309-2321.
6. Kisi, O. and J. Shiri, River suspended sediment estimation by climatic variables implication: Comparative study among soft computing techniques. Computers & Geosciences, 2012. **43**: p. 73-82.
7. Chang, C., et al., Appraisal of soft computing techniques in prediction of total bed material load in tropical rivers. Journal of earth system science, 2012. **121**: p. 125-133.
8. Khan, M.A., J. Stamm, and S. Haider, Assessment of soft computing techniques for the prediction of suspended sediment loads in rivers. Applied Sciences, 2021. **11**(18): p. 8290.
9. Guillet, G., et al., Fate of wastewater contaminants in rivers: Using conservative-tracer based transfer functions to assess reactive transport. Science of the Total Environment, 2019. **656**: p. 1250-1260.
10. Piasecki, M. and N.D. Katopodes, *Control of contaminant releases in rivers. I: Adjoint sensitivity analysis.* Journal of hydraulic engineering, 1997. **123**(6): p. 486-492.
11. Jamshidi, A., et al., Solving inverse problems of unknown contaminant source in groundwater-river integrated systems using a surrogate transport model based optimization. Water, 2020. **12**(9): p. 2415.
12. Kirkpatrick, J., et al., *Overcoming catastrophic forgetting in neural networks.* Proceedings of the national academy of sciences, 2017. **114**(13): p. 3521-3526.
13. Granata, F., et al., Machine learning algorithms for the forecasting of wastewater quality indicators. Water, 2017. **9**(2): p. 105.
14. Naseri-Rad, M., et al., INSIDE: An efficient guide for sustainable remediation practice in addressing contaminated soil and groundwater. Science of the Total Environment, 2020. **740**: p. 139879.
15. Mirghani, B.Y., et al., A parallel evolutionary strategy based simulation–optimization approach for solving groundwater source identification problems. Advances in Water Resources, 2009. **32**(9): p. 1373-1385.
16. Kargar, K., et al., Estimating longitudinal dispersion coefficient in natural streams using empirical models and machine learning algorithms. Engineering Applications of Computational Fluid Mechanics, 2020. **14**(1): p. 311-322.

17. Pourhosseini, F.A., K. Ebrahimi, and M.H. Omid, *Prediction of total dissolved solids, based on optimization of new hybrid SVM models.* Engineering Applications of Artificial Intelligence, 2023. **126**: p. 106780.
18. Zhang, H., et al., Proposing two novel hybrid intelligence models for forecasting copper price based on extreme learning machine and meta-heuristic algorithms. Resources Policy, 2021. **73**: p. 102195.
19. Gao, Z., et al., A novel multivariate time series prediction of crucial water quality parameters with Long Short-Term Memory (LSTM) networks. Journal of Contaminant Hydrology, 2023. **259**: p. 104262.
20. Sakaa, B., et al., Water quality index modeling using random forest and improved SMO algorithm for support vector machine in Saf-Saf river basin. Environmental Science and Pollution Research, 2022. **29**(32): p. 48491-48508.
21. Karim, T., et al., StackAMP: Stacking-Based Ensemble Classifier for Antimicrobial Peptide Identification. IEEE Transactions on Artificial Intelligence, 2024.
22. Ibrahim, D., *An overview of soft computing.* Procedia Computer Science, 2016. **102**: p. 34-38.