Journal of Applied Research in Water and Wastewater

E-ISSN: 2476-6283

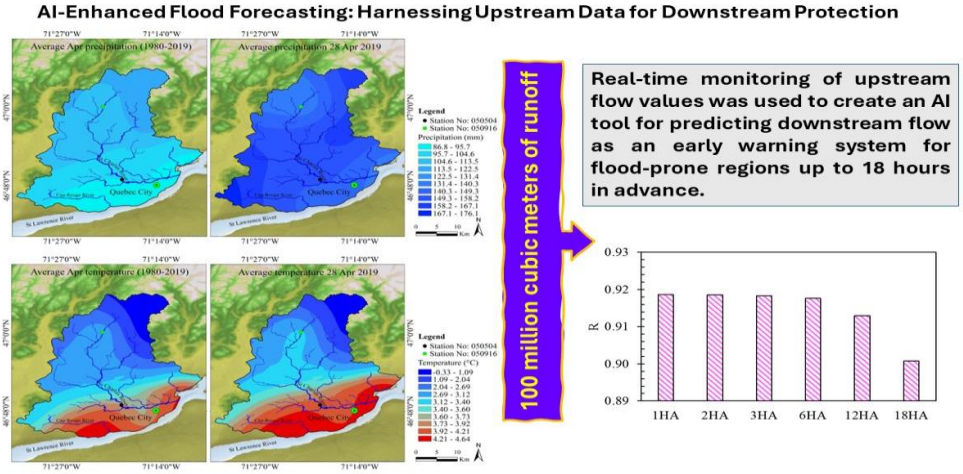
Journal homepage: https://arww.razi.ac.ir

AI-enhanced flood forecasting: Harnessing upstream data for downstream protection

Isa Ebtehaj[1,*], Hossein Bonakdari[2], Baram Gharabaghi[3]

[1]*Department of Soils and Agri-Food Engineering, Université Laval, Québec, Canada.*
[2]*Department. of Civil Engineering, University of Ottawa, Ottawa, Canada.*
[3]*School of Engineering, University of Guelph, Guelph, Canada.*

GRAPHICAL ABSTRACT



ARTICLE INFO

Article type:
Research Article

Article history:
Received 1 July 2023
Received in revised form 8 September 2023
Accepted 10 September 2023
Available online 11 September 2023

Keywords:
Artificial Intelligence
Flood prediction
Predictive analytics
Quebec
Water resource management

ABSTRACT

This research devised a cutting-edge artificial intelligence methodology to enhance flood forecasting in Quebec, Canada, an area frequently affected by floods. The core of this project was creating a novel artificial intelligence (AI) model (i.e., Generalized Structure of Group Method of Data Handling) dedicated to the early detection of potential flood events. Utilizing data from two key hydrometric stations, Saint-Charles and Huron, located within the region, the study aggregated data from 15-minute intervals into comprehensive hourly averages. An initial analysis sought to understand the relationship between river flow rates and the environmental factors of temperature and precipitation upstream and downstream. The investigation uncovered intricate relationships among these factors, presenting challenges in accurately predicting floods. To address this, a specialized AI model was developed to translate the flow data from the Huron station to predict potential flooding at the Saint-Charles station. This model, leveraging 48-hour lag data from upstream, was designed to forecast flood events at the Saint-Charles station with lead times ranging from one to eighteen hours. The model demonstrated significant predictive accuracy, with a correlation coefficient surpassing 0.9. Consequently, this innovative AI model emerges as a promising tool for improving Quebec's flood prediction and early-warning systems.

1. Introduction

The failure to adapt or mitigate climate change represents the most significant challenge facing communities worldwide in the future decade. The dynamics of vital hydro-climatic variables worldwide have experienced remarkable changes due to climate change. Floods, as one of the critical hydro-climatic variables, are the most frequently happening natural hazard to the environment, infrastructure, property, economy, and life around the globe, especially in Canada (Ebtehaj and Bonakdari, 2023). According to the damages caused by floods, It stands as one of the most lethal catastrophes globally, so it is ranked third after the earthquake and tsunami (Zaji *et al*., 2018). Some

*Corresponding author Email: isa.ebtehaj.1@ulaval.ca

Cite this article: I. Ebtehaj, H. Bonakdari, B. Gharabaghi, AI-enhanced flood forecasting: Harnessing upstream data for downstream protection, *Journal of Applied Research in Water and Wastewater*, 10 (2), 2023, 119-132. doi: https://doi.org/10.22126/arww.2024.10226.1322

prominent causes of flooding can be expressed as the runoff from melting snow in spring, precipitation from storms, barriers formed by nature, inundation along coastlines, waterlogging in urban areas, and failure of flood management structures, and groundwater (Shrubsole *et al.*, 1993; Letessier *et al.,* 2023; Ebtehaj *et al.,* 2023a). According to the large variety of hydrological features, weather, and landscape in Canada, floods are triggered through three main mechanisms such as rain on snow, heavy rainfall, and spring snowmelt (Zahmatkesh *et al.,* 2019), which is the most prominent causes of floods in Canada (Shrubsole *et al.,* 2003). Floods occur most along with large river systems in the spring, when maximum flow rates are predominantly determined by the volume of runoff resulting from snowmelt and rain, yet can also happen in the summer (Hildebrandt, 2013), characterized by sudden flooding in city environments due to heavy rainfall over brief periods.

Several significant floods have occurred in Canada in previous decades. One of the oldest ones is the Fort Calgary flood in 1879, with a maximum flow rate of 2265 m^3/s corresponding to a 1:200 year event for Bow River. The flood in Southern Alberta resulted in the largest evacuation of a natural disaster in Alberta's history, leading to around 100,000 Albertans leaving their homes in June 2013. An average of 75 to 150 mm of rainfall over three days was the main reason for this catastrophic event, with four deaths. The Elbow and Bow rivers in Calgary were flooded, and 3,000 buildings overflowed, affecting over 4,000 businesses and compelling approximately 75,000 residents to evacuate the city. This event cost $2,715,742,000 (Canadian Disaster Database (CDD)). This flooding represents the highest estimated financial impact of any flood in Canada. Southern Manitoba (June 2014) and Southern Alberta and Saskatchewan (June 2010) floods with $1,164,679,000 and $1,031,670,000 are ranked second and third (respectively) with the highest estimated total costs.

Floods can occur with a wide spatiotemporal variation in Canada, with different severity and damage based on the mainstream involved, location, and watershed size. An improved early flood detection system and mitigation measures are crucial for decision-makers in watershed management. Accurate and reliable flood anticipating can decrease some of the adverse effects of flood events and provide reliable and accurate predictions with suitable lead time. Hydrologic models serve as analytical instruments for systems, capable of mimicking hydrological behaviors to reanalyze the historical incident and/or forecast the responses of the basin to the future hydrological process.

A conceptual model and a physically based model can be classified according to their underlying equations, while lumped and distributed catchment areas can be classified according to how the catchment area is represented (Sahraei, Asadzadeh, Unduche, 2020). A variety of physically-based hydrological models (PBHM) are used for water management, such as WEAP (Gao, Christensen, Li, 2017), MIKE Basin (López *et al.,* 2020), the Hydrologic Engineering Center's Hydrologic Modeling System (HEC-HMS) (Ramly *et al.,* 2020), and the Soil and Water Assessment Tool (SWAT) (Li *et al.,* 2018). Various physically-based hydrological models (PBHMs) are available for water management in the watershed, including SWAT (Li *et al.,* 2018), MIKE Basin (López *et al.,* 2020), HEC-HMS (Ramly *et al.*, 2020), and WEAP (Gao *et al.,* 2017). Nonetheless, while PBHM models have shown remarkable capabilities in predicting different types of flooding, they require extensive computational resources and a variety of hydro-meteorological and geomorphological data, such as soil characteristics, land use, vegetation, and slope. The necessity for numerous input variables, each prone to measurement errors, can accumulate, adversely affecting the outcomes of the models (Kollet and Maxwell, 2006). Furthermore, missing data in any of the input variables can significantly limit the practicality of PBHMs (Soltani *et al.,* 2021). Thus, employing a PBHM for flood prediction is notably complex, demanding specialized knowledge and a comprehensive understanding of hydrological variables, which presents significant challenges (Kim *et al.*, 2015).

Successful application of Artificial Intelligence (AI) in solving complex nonlinear problems persuades decision-makers and scholars toward AI techniques to model the nonlinear behavior of flood events and their characteristics in different stages of floods utilizing past data without understanding the physical mechanics of flooding. AI models are applied to induce patterns and regularities through fast modeling with high performance (Ebtehaj and Bonakdari, 2016; Khozani, Bonakdari, Ebtehaj, 2017; Ebtehaj, Bonakdari, Zaji, 2018; Sihag *et al.*, 2019; Safari *et al.*, 2019), as well as more straightforward implementation, less complexity, and lower computation compared to PBHM (Mekanik *et al.,* 2013). Application of the AI models in two past decades showed the pertinence of these models in flood forecasting with an admissible rate of outperforming classical techniques (Walton *et al.*, 2019). Different AI approaches, including artificial neural network, neuro-fuzzy (Mosavi and Edalatifar, 2019), extreme learning machine (Taherei Ghazvinei *et al.*, 2018) support vector regression (Gizaw and Gan, 2016), and deep learning (Puttinaovarat and Horkaew 2020) were identified as effective instruments for predicting short- and long-term floods.

The innovation of this study unfolds in three distinct dimensions. Initially, it introduces a novel AI methodology, specifically the Generalized Structure of the Group Method of Data Handling (GSGMDH), tailored for early-warning flood forecasting within Quebec province, Canada. Secondly, an exhaustive literature review reveals the absence of prior flood forecasting research employing comparative analysis across various potential input combinations. This investigation rigorously evaluates an extensive array of input configurations, ranging from two to 48 variables, utilizing the GSGMDH model's automatic capabilities. Lastly, whereas conventional approaches predominantly rely on recorded discharge data at the target station for future flood predictions, this research diverges by leveraging discharge data from upstream locations to enhance the accuracy of downstream flood forecasts. The 15-minute flow rates at two hydrometric stations, including Hurons and Saint-Charles, are collected to develop it. Using the hourly averaging of this data, six different AI models with one to eighteen lead times are developed to forecast the flow rate at Saint-Charles, located upstream of Quebec City, using the historical flow rate at Hurons station. Decision-makers need a relationship that can be used by an AI with the most minor parameters in practical tasks. This system can be an alarm system in which decision-makers are alerted to floods downstream when the upstream flow rate reaches a particular value.

2. Materials and methods

2.1. Importance of flood forecasting in Quebec province

Flooding is Canada's most common natural calamity. According to the CDD (https://cdd.publicsafety.gc.ca), 35 flood disasters happened in Quebec, which is more than 10% of floods that occurred in Canada (i.e., 309 floods in all provinces) from 1900 to 2013. The number of floods in Canada and Quebec is more than three times the of wildfires as the next most usual natural hazard.

In the last century, Quebec has witnessed significant flooding events. The Saguenay flood on July 19, 1996, occurred due to 290 mm of rainfall in less than 36 hours. As a result of this flood, thousands of bridges, roads, and homes were washed out, and at least led to 10 deaths. More than 15,000 people had to evacuate their houses. The projected overall expense of the Saguenay flood was 300 million CAD (CDD, https://cdd.publicsafety.gc.ca). In the 2019 Quebec flood, 51 municipalities with 6681 residences were flooded in five zones due to the submerged roads and landslides, which led to over 13500 disaster victims ("Inondations: plus de 10 000 personnes évacuées". *La Presse.* April 30, 2019). The type of this flood was rain in the region and snowmelt that caused lakes and rivers to overflow (http://floodlist.com/america/canada-floods-quebec-april-2019). The approximated expenses for the flood disasters in the Quebec province of Canada from 1974 to 2014 were collected from the CDD. The summation of all Available estimated costs of flood disasters in Quebec is more than 840 million CAD. Fig. 1 provides the historical record of flood events in Canada from 1900-2017. Based on the figure, Ontario experienced the highest number of flood events, with 49 flood disasters, while the lowest one was related to Saskatchewan, with two flood disasters. With 27 floods through these years, Quebec has more than 14% of all floods in Canada. Besides, the cumulative total damage related to Quebec is 3218 million CAD, more than 12% of all cumulative total damage in Canada.

Fig. 2 indicates the dominant flood types for each Canadian province. Due to this Fig., none of them are the prevalent flood types in the country. Due to this Fig., heavy rainfall is the leading cause of flooding in most Canadian provinces, including Quebec. Besides, snowmelts, ice jams, and riverine flooding are also causes of floods in Quebec province. Due to CDD, the leading cause of flooding in Quebec is heavy rainfall, which is the cause of more than 52% of all floods in this province. The riverine flooding, ice jams, and snowmelts are ranked second to fourth with more than 19%, 16%, and 11%, respectively.

According to the provided explanations, timely dissemination and precise flood forecasting in the Quebec province is essential to the decision-makers, policy, and the public. Precise flood prediction can diminish the social and economic impacts of floods on communities, minimize infrastructure damage, and lead to measures that can improve ecological conditions.

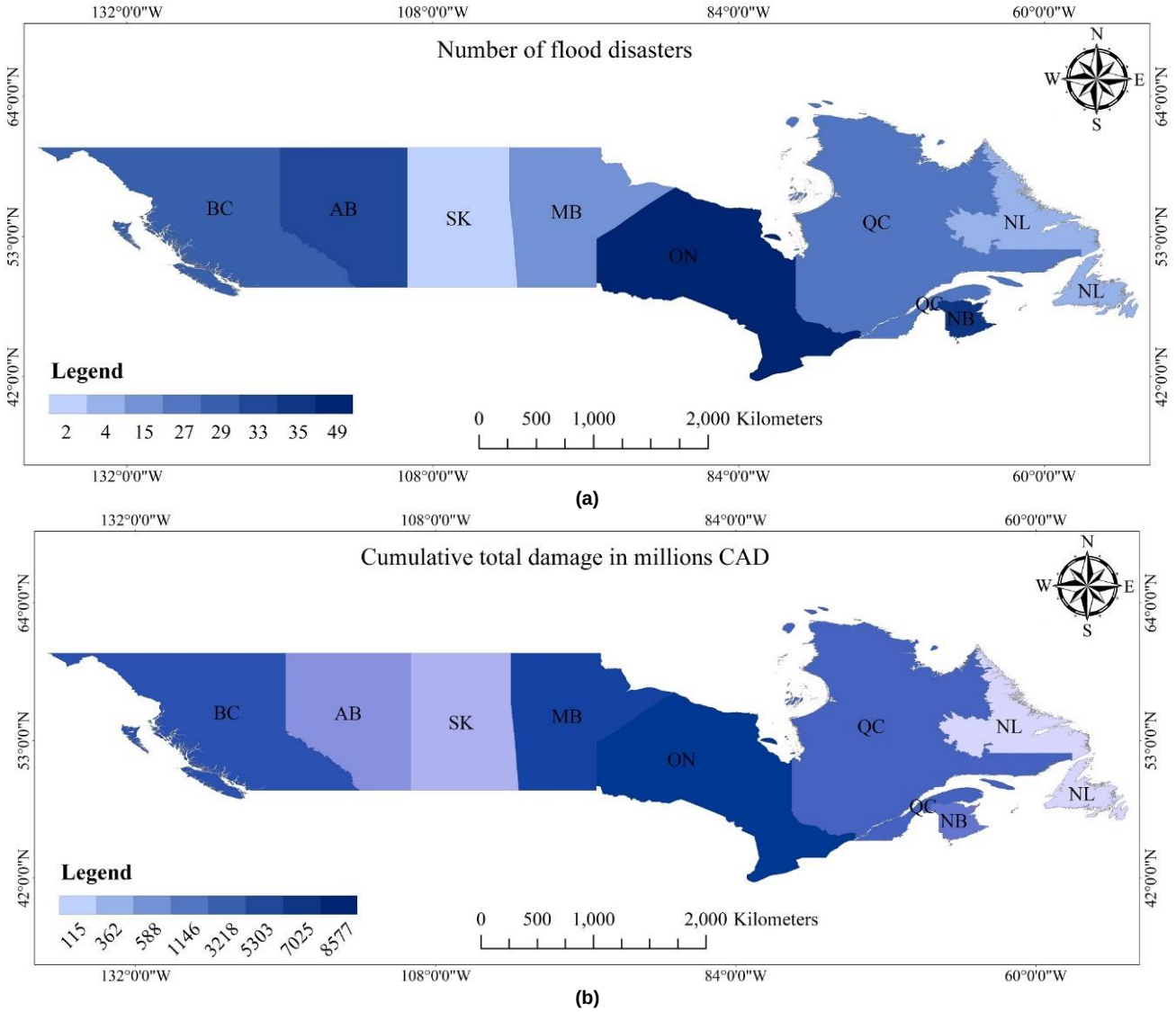


Fig. 1. History of the flood incidents in Canada from 1900-2017 (McGrath, Stefanakis, Nastev, 2014; http://www.museedufjord.com/inondations/manitoba_eng/tumultuous): (a) Number of flood disasters; (b) Cumulative total damage in millions CAD.

2.2. Region of study

The location of the study area is depicted in Fig. 3. According to this Fig., it is situated in Quebec province, Canada. Around Quebec City, four different sub-catchments can be envisaged within the Saint-Lawrence watershed. The study area of the current study is located at sub-basin "a". As shown in Fig. 3, two hydrometric stations known as 050916 and 050904 are employed, and they are placed upstream and downstream (respectively) of the Saint-Charles River. The 050916 and 050904 are known as Hurons and Saint-Charles, respectively. The Saint-Charles station is located 0.8 km upstream from Lorette, while the Hurons are located at the Crawford Street Bridge in Stoneham.

The Hydrographic region of both stations is Saint-Laurent northwest. All the flow rate data (m^3/s) are collected from the Ministry of the Environment, the Fight against Climate Change, Wildlife and Parks (https://www.cehq.gouv.qc.ca). The data collected from both stations were recorded every 15 minutes, and their hourly average was established to produce an hourly river flow dataset. The collected data cover twelve years of observation from April 08, 2008, to December 09, 2020. It should be noted that these data include missing ones from November to April. The desired delays were defined to deal with missing data, and the rows with missing data were eliminated from the dataset. The number of all samples was 74707, randomly categorized into two groups: train and test. A total of 52,295 samples were chosen for model calibration., and the rest were applied to validate it. The maximum flow rates (m^3/s) for 050916 and 050904 stations are103.48 and 97.48, while the mean value of the flow rates (m^3/s) for 050916 and 050904 stations are 3.25 and 10.98, respectively. The ratios of the maximum to the mean of flow at 050916 and 050904 stations are more than 30 and 9 times, respectively. The training and testing data distributions for both stations are provided in Fig. 3.

2.3. Generalized structure of group method of data handling

Theoretically, understanding the explicit mathematical connections between input variables and their corresponding outcomes is essential to model a system effectively. It is challenging to extract explicit modeling, and these connections remain elusive in many systems. Under these circumstances, approaches that utilize data to make calculations based on input and output records are taken into account. Such strategies are highly effective in discerning the complexities of nonlinear systems. This study presents the early flood warning system model using the generalized structure of the group method of data handling (GSGMDH). This model allows early flood forecasting without the need to estimate complex parameters and only uses the flow rate upstream in a short time so that it can be used effectively and efficiently by water managers and planners.

The GMDH algorithm stands as a neural network framework featuring layers for input, intermediary processing, and output. Various studies on the application of GMDH show that the use of approximation and optimization combinations in the structure of this method produces more accurate results in predicting the physical behavior of phenomena (Azimi et al. 2018). The GMDH method is a self-organizing technique where models progressively develop a more complex architecture by assessing their effectiveness across a series with multiple inputs and corresponding output. This method gradually develops the model structure during its performance evaluation due to its internal self-organizing approach; Hence, it is smarter than other algorithms. The core concept of this approach involves building an analytical function within a feedforward neural network, utilizing a quadratic transfer function. The coefficients for this function are determined through the least squares approach.

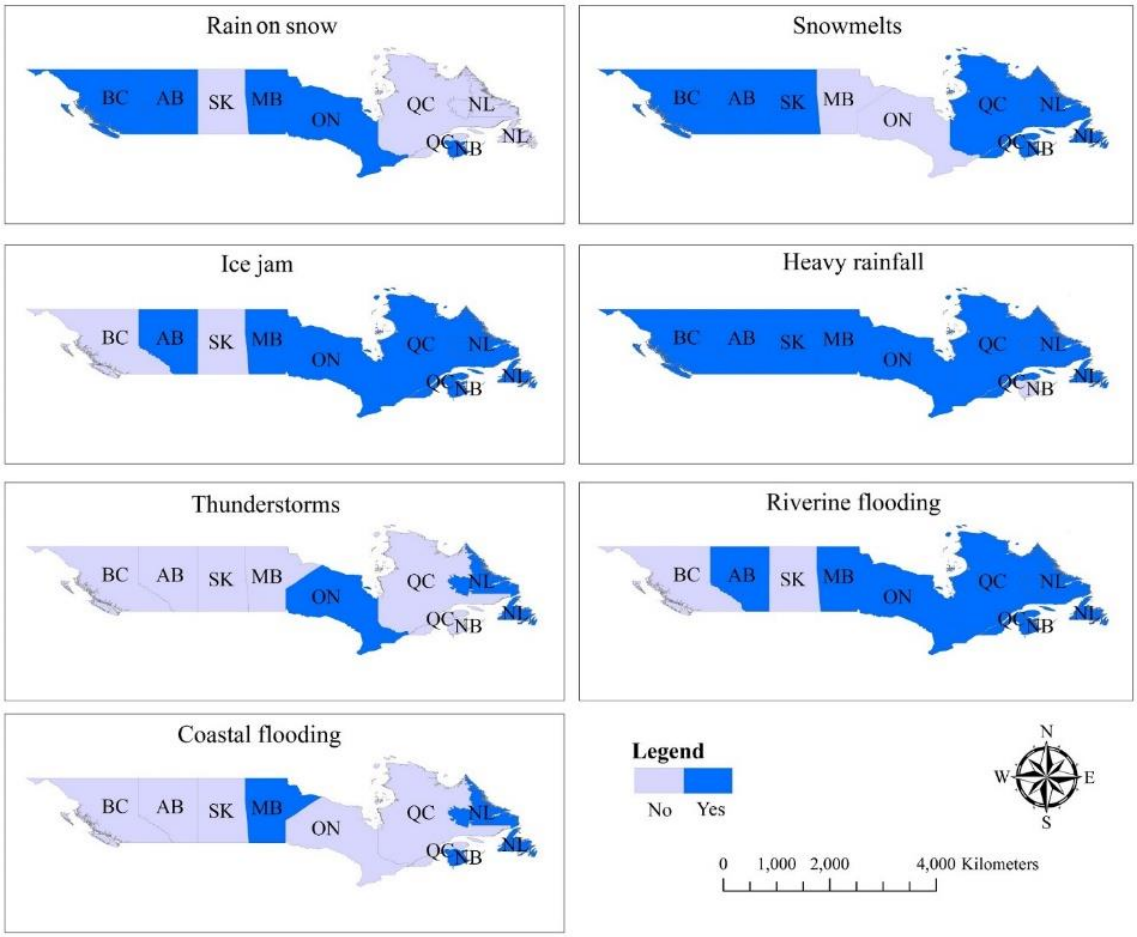
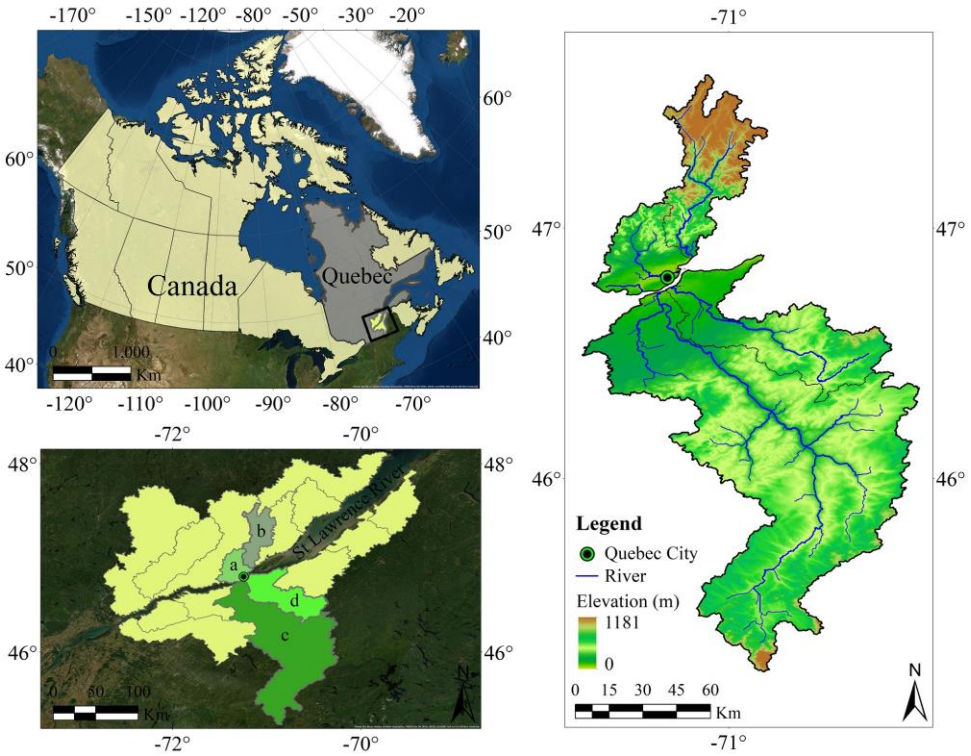


Fig. 2. Flood disasters in Canada by type in each province (Data from Zahmatkesh et al. 2019).

.



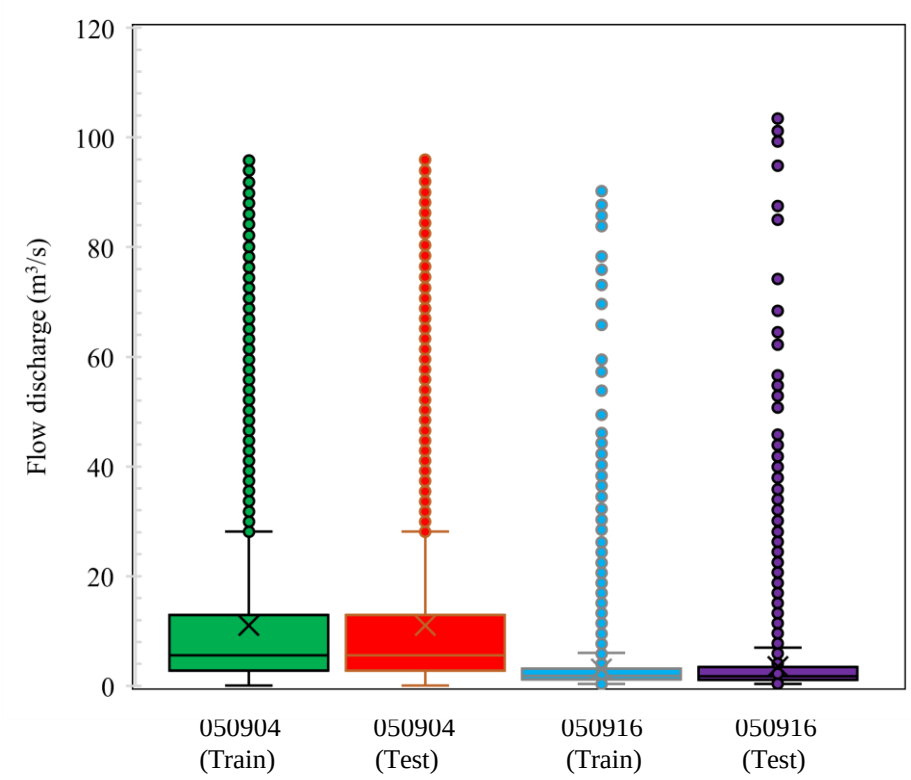


Fig. 3. The geographical positioning of the area under study and data distribution.

The architecture of this model includes multiple neurons in every layer, each formed by linking a quadratic polynomial to two distinct inputs. Similarly, neurons formed in preceding layers are introduced into subsequent layers as fresh inputs, facilitating the generation of additional neurons. Within this network, the number of neurons per layer matches the total of binary combinations derived from network input variables. Each neuron is characterized by two input variables and a single output variable. Essentially, the main objective of this technique is to amalgamate quadratic polynomials from every neuron to craft the approximate function ŷ, which forecasts the output for a given set of inputs with minimal deviation from the real output y. The quadratic forms of the Eqs. in the GMDH are as follows:

$$\hat{y} = f(x_1, x_2) = I_0 + I_1 x_1 + I_2 x_2 + I_3 x_1^2 + I_4 x_2^2 + I_5 x_1 x_2 \quad (1)$$

where I = {I_0, I_1, I_2, I_3, I_4, I_5} is the set of unknown coefficients optimized through the training phase, and x_1 and x_2 are the input neurons of the desired Eq.

An example of the GMDH with four input neurons is indicated in Fig. 4. In this Fig., x_1, x_2, x_3, and x_4 are the input variables, while the x_{LN} denotes the N^{th} generated neuron in the L^{th} layer. For example, the x_{21} is the first neuron of the second layer. Due to this Fig., five different neurons (x_{11}, x_{12}, x_{13} in layer one and x_{21}, x_{22} in layer 2) were generated to map the input variables (i.e., x_1, x_2, x_3, x_4) to output variable (y). All neurons were generated using the neurons in the adjacent previous layer. For example, x_{11} was produced using x_1 and x_4, and x_{21} was created using x_{12} and x_{13}. The adjacent layer for the x_{11} is the input layer, while it is layer 1 for the x_{21}. Another point that should be considered in this Fig. is that each new neuron could be generated from only two neurons. To connect the input variables to output variable, as indicated in Fig. 4, six different quadratic equations in the form of Eq. 1 were generated. The number of combinations evaluated by GMDH to achieve the best structure for every layer is determined as follows:

$$NNEL = \binom{n}{2} = \frac{n(n-1)}{2} \quad (2)$$

where NNEL is the number of neurons in each layer, and n is the maximum input in each layer. Considering the provided example in Fig. 4, the NNEL in layers 1 and 2, as well as the output layer, is 6, 3, and 2, respectively.

According to the provided explanations, the GMDH has some advantages, including (i) Automatically determining influential input variables among other variables, (ii) Automatic identification of the model's architecture, including layers' number and inputs of each layer, and (iii) Consider the accuracy and simplicity of the model simultaneously to prevent over-fitting using the Akaike Information Criterion (Zeynoddin et al. 2019). In addition to the advantages provided by this method, GMDH has some drawbacks that significantly impact the modeling accuracy. The mentioned limitations include quadratic polynomials with only two input neurons and choosing the input neurons solely from the neighboring layer. The first one may affect the failure to model complex nonlinear problems, while the second results in highly complicated models with many newly generated neurons. To overcome these drawbacks, an updated variant of the GMDH, referred to as the generalized structure of GMDH (GSGMDH), is developed in the current study. In this method, the inputs of each new neuron could be two and/or three chosen from either the neighboring or non-neighboring layer with second and/or third-order polynomials.

The results of GSGMDH-based modeling of the provided problem in Fig. 4 are presented in Fig. 5. Due to this Fig., the x_{11} has three inputs, including x_1, x_3, and x_4, while the x_{21} has only two inputs including x_2 and x_{11}. It should not be that the inputs of x_{21} are chosen from both neighboring (x_{11}) and non- neighboring (x_2) layers. Besides, the inputs of the output layer are also chosen from both adjacent (x_1 and x_{11}) and non-adjacent (x_{21}) layers. Therefore, the generated new neurons were reduced from 6 in GMDH to 3 in the GSGMDH.

2.4. Performance statistics for model evaluation

Four different statistical metrics are utilized to evaluate the effectiveness of the GSGMDH in early-warning flood prediction. They are divided into three primary categories, including correlation-based metric as correlation coefficient (R), absolute metrics including normalized root mean square error (NRMSE) and mean absolute error (MAE), as well as relative one including root, mean square relative error (RMSRE). Based on the literature studies, the combination of indices is sufficient for assessing the model's performance (Ebtehaj et al. 2023b; Lotfi et al. 2019; 2020). The mathematical definition of the R, MAE, NRMSE, and RMSRE are as follows:

$$R = \frac{\sum_{i=1}^{N}(A_i - \bar{A})(M_i - \bar{M})}{\sqrt{\sum_{i=1}^{N}(A_i - \bar{A})^2}\sqrt{\sum_{i=1}^{N}(M_i - \bar{M})}} \quad (3)$$

$$MAE = \frac{\sum_{i=1}^{N}|A_i - M_i|}{N} \quad (4)$$

$$NRMSE = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(A_i - M_i)^2}}{\sum_{i=1}^{N} A_i} \quad (5)$$

$$RMSRE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{A_i - M_i}{A_i}\right)^2} \quad (6)$$

where N stands for the samples' number, A_i signifies the i^{th} sample, $\overline{A}$ represents the mean of actual samples, while M_i and $\overline{M}$ are the i^{th} sample and mean of modeled samples, respectively.

3. Results and discussion

Fig. 6 demonstrates the peak flow distribution through different years and months. The defined classes in this Fig., including C1, C2, and C3, are related to flow rate (m^3/s) in the range of [80, 100], [60, 80), and [30, 60), respectively. The maximum percentage of the peak flow rate is related to 2019, 2018, and 2017 for C1, C2, and C3 (respectively), with 46.61%, 21.51%, and 14.04%, respectively. Indeed, the strongest floods with a flow rate of more than 80 m^3/s happened in 2019. As the peak flow decreases (i.e., less than 80 m^3/s), the percentage is distributed among the different years so that the maximum percentage recorded for C2 and C3 is less than half the percentage obtained for C1.

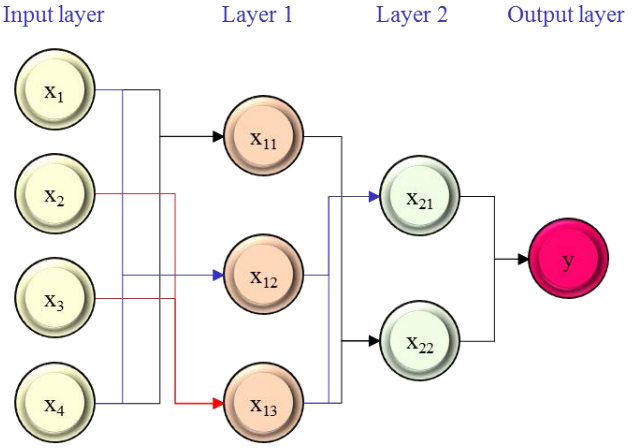


Fig. 4. An example of the GMDH with four input variables.

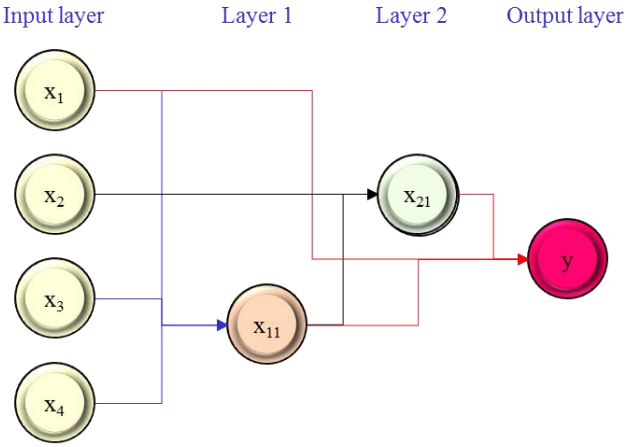
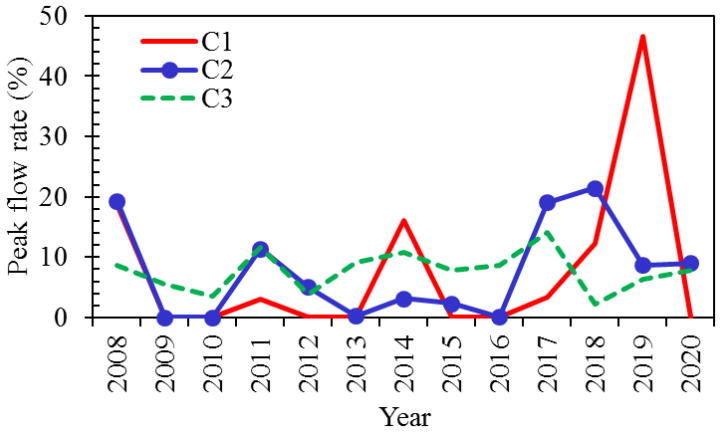
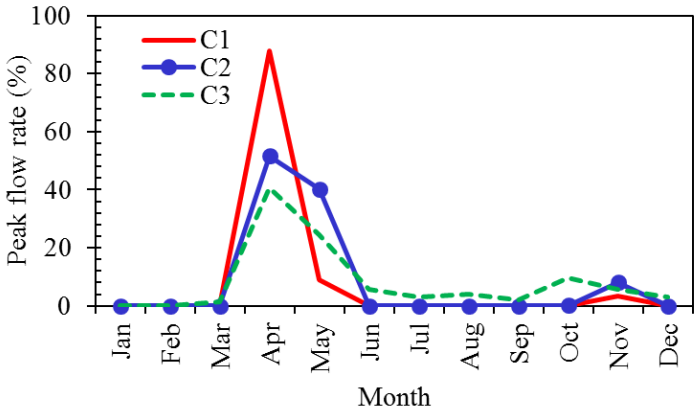


Fig. 5. An example of the GSGMDH with four input variables.



(a)



(b)

Fig. 6. Peak flow distribution through different years and months.

The distribution of the peak flow rates also shows that the maximum flow rate percentage at all classes is related to April, with 87.71%, 51.6%, and 40.8% for C1, C2, and C3, respectively. May is also ranked second in all classes, with 8.9%, 40.22%, and 24.48% for C1, C2, and C3, respectively. For a flow rate of more than 60 m^3/s (i.e., C3 and C2), more than 90% of all peak flow occurred in April and May. The main reason for the flood in these two months could be due to the decrease in temperature, which also led to the snow melting and heavy rainfall in these two months. Two floods in 2008 and 2019, along with temperature and precipitation values related to different stations, are evaluated to investigate the reason stated. Fig. 7 shows the value of the flow rate (m^3/s), precipitation (mm), and temperature (°C) from 04/08/2008 to 05/07/2008. The maximum value of the flow rate at station 50904 is 97.48 m3/s, recorded at 6 p.m. on 4/30/2008. According to Fig. 3, Station 509016 is upstream, while Station 50904 is downstream. To check the effect of flow rate on the upstream station (i.e., 509016) and meteorological parameters, including precipitation and temperature at both stations, on the flood that occurred at station 50904, the maximum value of the mentioned parameters must be explored. The maximum flow rate value at station 50916 is 36.785 m^3/s, recorded at 7 p.m. on 4/29/2008. Indeed, the difference between the maximum value of the flow rate of the upstream and downstream stations is 25 h.

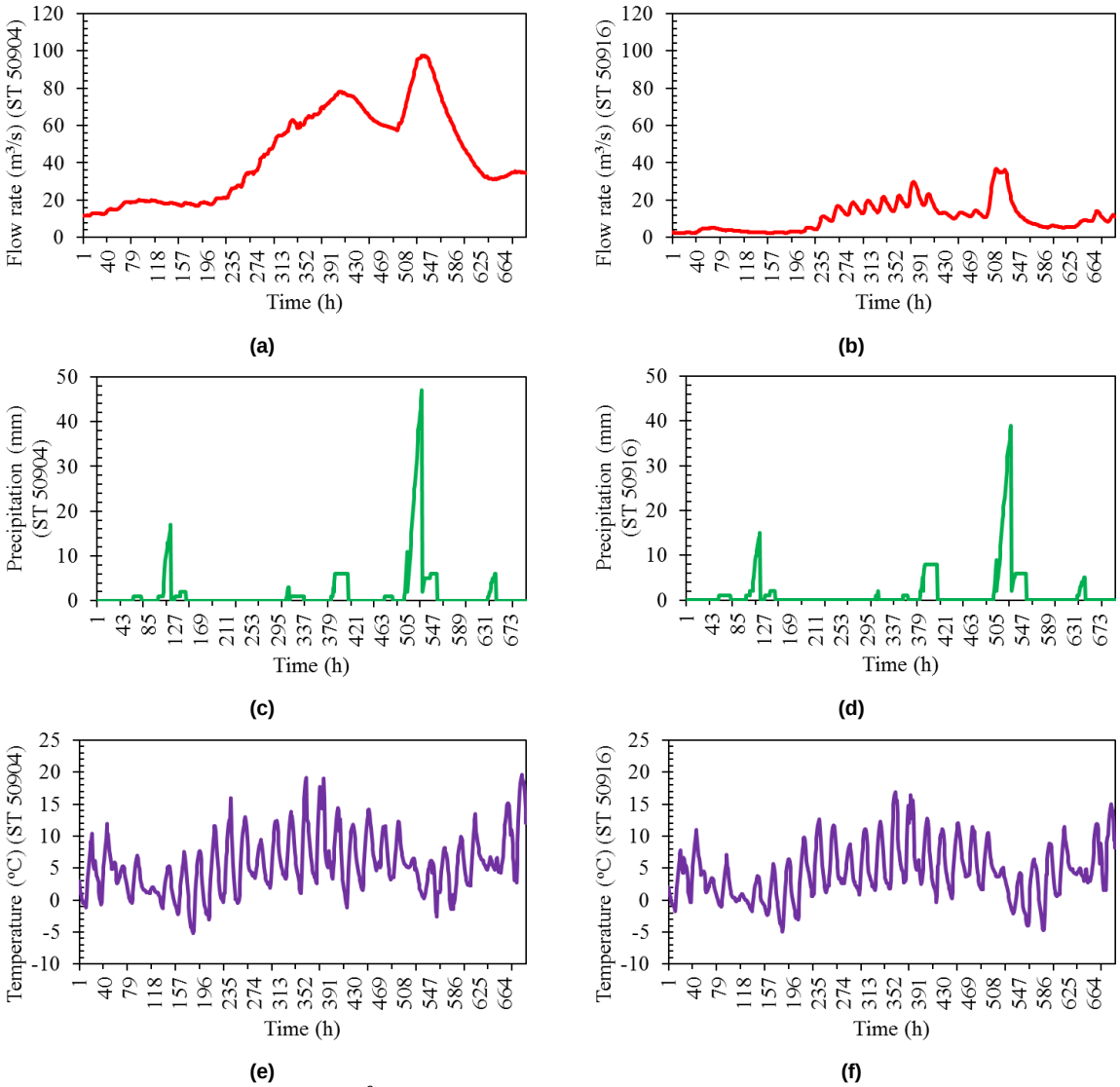


Fig. 7. The value of the flow rate (m^3/s), precipitation (mm), and temperature (°C) from 04/08/2008 to 05/07/2008.

Given that the difference in peak flow rates between the two stations is significant, other factors may influence the floods at the downstream station. The maximum precipitation at stations 50916 and 50904 was 38 mm and 43mm (respectively) recorded on 24 and 23 of

4/29/2008, seven and six hours before the flood at the downstream station. Therefore, it can be said that in addition to the flood upstream, the precipitation six hours ago could also be one of the reasons for the flood in the downstream station. Besides, the increase in temperature in the days before 4/30/2008 at both upstream and downstream stations can be another factor for the occurrence of floods at the downstream station, so the maximum temperature recorded in the upstream and downstream stations are 19.056°C and 16.433°C (respectively), which were recorded seven days before the flood occurred at downstream station. Temperature increases lead to snowmelt. The results of Fig. 7 confirmed that snowmelt and heavy rainfall are the most well-known reasons for floods in Canada, as indicated in Fig. 2 and also reported by Shrubsole et al. (2003) and Zahmatkesh et al. (2019).

Fig. 8 illustrates the value of the flow rate (m^3/s), precipitation (mm), and temperature (°C) from 04/18/2019 to 05/17/2019. The maximum flow rate value at station 50904 is 96.18 m3/s, recorded at 9 a.m. on 4/28/2019, while it is 47.52 m3/s for station 50916, recorded at 11 a.m. on 4/27/2019. Indeed, it takes about 22 hours for the peak flow rate to reach the effect from the upstream station (i.e., 50916) to the downstream station (i.e., 50904), compared to 25 hours for the sample shown in Fig. 7. The difference between the peak flow rates of the two stations is significant, which can be due to the effects of other parameters, including precipitation and temperature, as well as geological characteristics. To explore the impact of the upstream station on the downstream station further, the precipitation and temperature at both stations through the mentioned dates are also checked. The maximum precipitation at stations 50916 and 50904 was 43 mm and 41 mm (respectively) recorded on 22 of 4/27/2019, 15 hours less before the flood at the downstream station. Indeed, the maximum precipitation at the upstream station occurred within less than one day of the flooding time at the downstream station, while for the example shown in Fig. 7, this time was about 6 hours. Precipitation was reported to be zero at both stations six hours before the floods. The temperature at the upstream station had risen from 26.35°C to 47.5 °C (80% increase) six days before the floods began at the downstream station. While the highest value of this parameter was recorded in the downstream station on 4/21/2019 (T = 14.46 °C). It can be said that a significant increase in temperature at the upstream station has led to snow melting and has been one of the significant factors in the flood occurrence downstream.

Fig. 9 shows the map of the precipitation and temperature from April through 1980-2019 and its value on April 28, 2019. Due to this Fig., the average precipitation (mm) of April through 1980-2019 is in the range of [90.75, 119.8] while it is [138, 171] for April 28, 2019. Indeed, the minimum and maximum values of precipitation were increased by more than 50% and 40%, respectively. It could be seen that for average precipitation of April through 1980-2019, the lowest precipitation was recorded around Quebec City, while for April 28, 2019, it was almost 150 mm, which is 8% higher than the lowest one and 12% lower than the highest one in this date. It should be noted that the average volume of precipitation in April through 1980-2019 and April 28, 2019, are more than 75 and 100 million cubic meters, respectively.

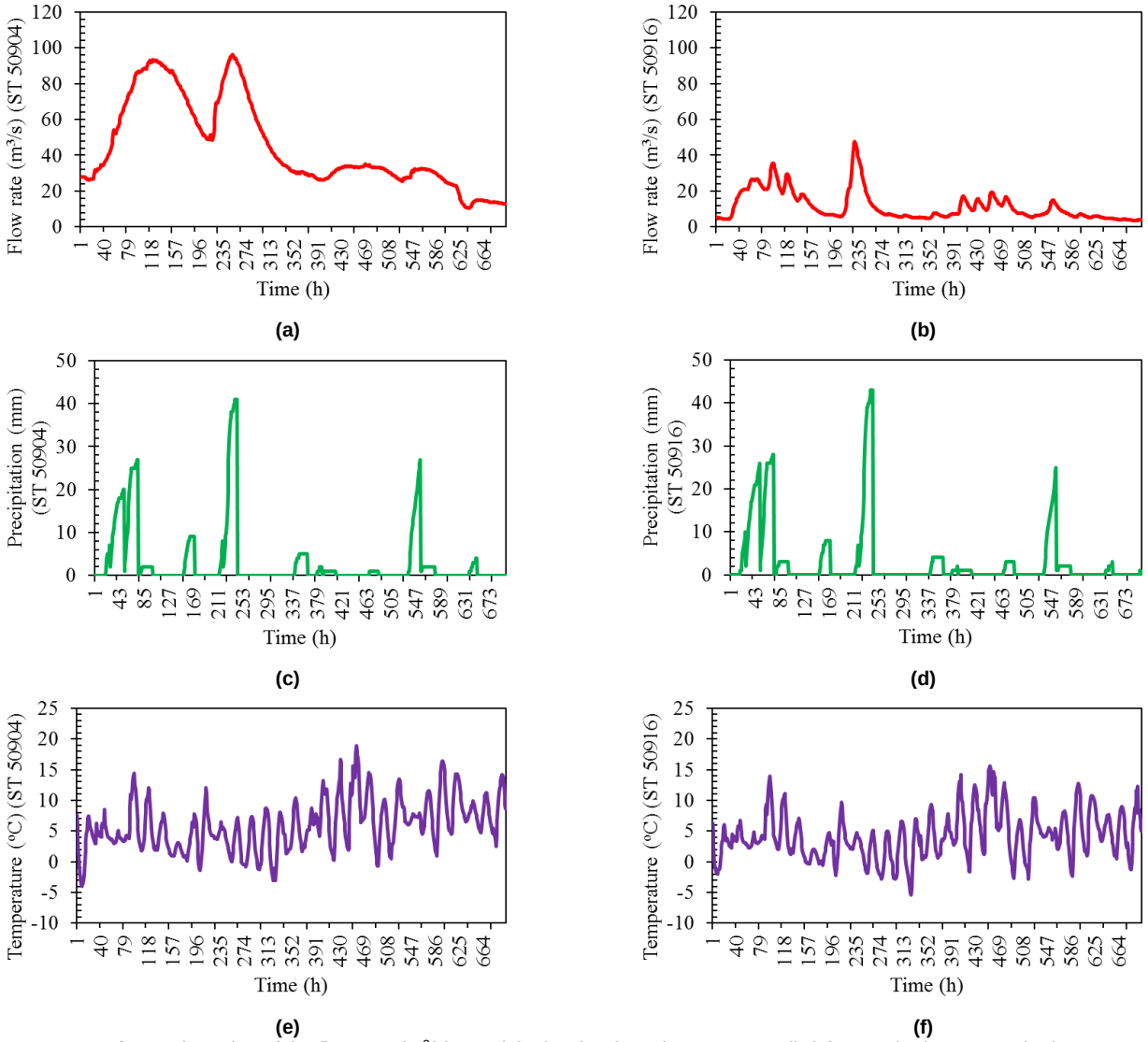
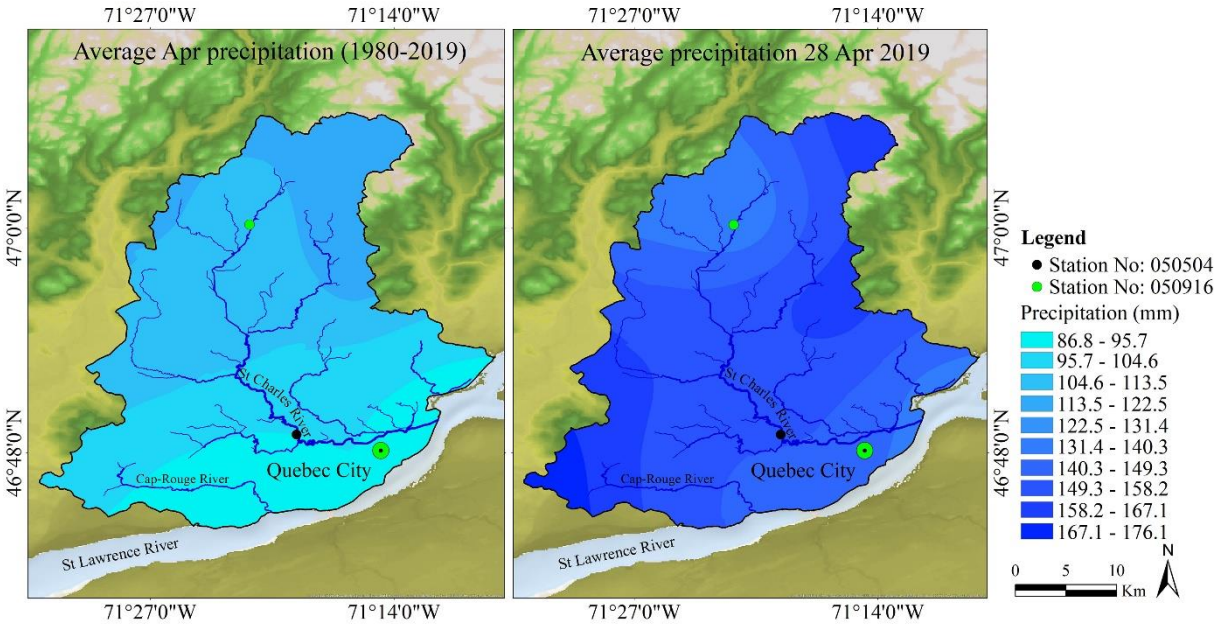
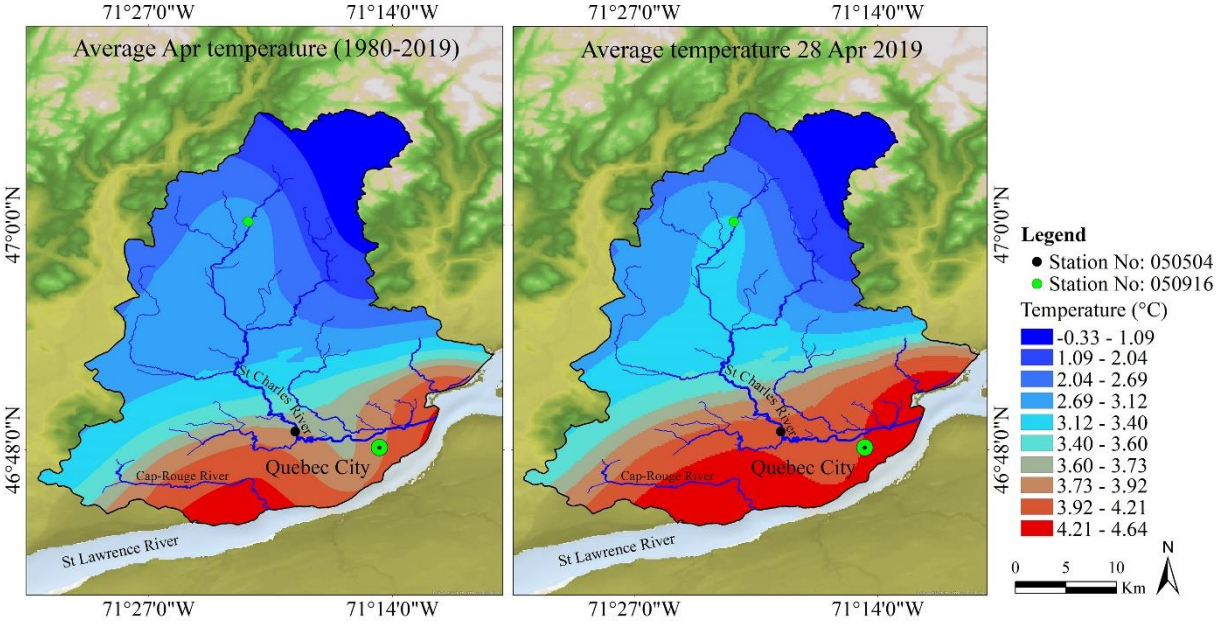


Fig. 8. The value of the flow rate (m^3/s), precipitation (mm), and temperature (°C) from 04/18/2019 to 05/17/2019.



(a)

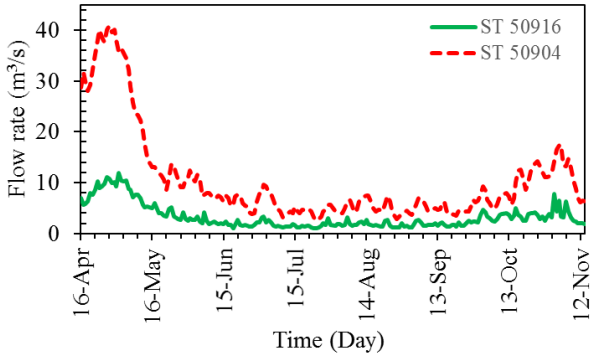


(b)

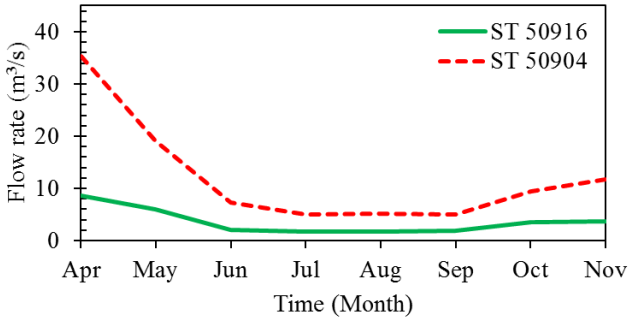
Fig. 9. The map of (a) the precipitation and (b) temperature through 1980-2019 and its value on April 28, 2019.

In addition to precipitation, the temperature distribution from April through 1980-2019 and its value on April 28, 2019, is also provided in Fig. 9. Due to this Fig., the average temperature (ºC) of April through 1980-2019 is in the range of [-1, 4] while it is [0, 4] for April 28, 2019. Indeed, the maximum value of the historical temperature at April and April 28, 2019, is almost equal. According to Fig., the temperature on April 28, 2019, is 1 ºC greater than the historical value of this temperature in April.

Fig. 10 shows the flow rate distribution for each day and month. It should be noted that days started from April 16 to November 14. The flow rate on other days is not available. Besides, the provided months are in the range of April and November, April starts on the 16th day. According to this Fig., the maximum flow rates were recorded on May 2 and April 28 for stations 50916 and 50904, respectively. Meanwhile, the maximum average flow rate at both stations was recorded in April.



(a)



(b)

Fig. 10. The distribution of the flow rate at each (a) day and (b) month.

Given the explanations provided in Figs. 7 and 8, it could be concluded that there is a very complex relationship between flow rates at upstream and downstream stations that is not easily predictable by using precipitation and temperature measurements at different stations. Besides, the results in Fig. 9 show that using long-term data, commonly used by decision-makers, may not be the right choice for infrastructure planning. Therefore, it is required to develop a supervised AI-based approach to map the flow rate downstream to the upstream one using daily recorded samples.

Given the explanations provided in Figs. 7 and 8, it could be concluded that there is a very complex relationship between flow rates at upstream and downstream stations that is not easily predictable by using precipitation and temperature measurements at different stations. Therefore, it is required to develop a supervised machine learning-based approach to map the flow rate downstream to the upstream one. For this purpose, the flow rate upstream is used as 48 different inputs, including one to 48 days before (i.e., 48 delays) to estimate the flow rate downstream, as follows:

1HA:

$$FL_d = f(FL_u(t-1), FL_u(t-2), ..., FL_u(t-48)) \quad (7)$$

where FL_d and FL_u are the flow rates downstream and upstream, respectively, and IHA means one hour ahead.

The use of 48 different inputs in a modeling process is too much. Fortunately, due to the ability of the GSGMDH method to find the number of effective inputs, only the parameters are considered as inputs that exert the greatest influence on the effectiveness of the model. The AIC index (Ebtehaj et al. 2020) is used to select the best model structure. This index considered the accuracy and simplicity of the model simultaneously to prevent a highly complex model that reduces its generalizability and leads to overfitting for unseen data. Eq. 7 is presented for a situation where our goal is to predict one hour ahead. In early-warning flood prediction, the use of just one hour is not significant, and more time needs to be evaluated.

Therefore, in this study, in addition to one-hour ahead prediction, two, three, six, twelve, and eighteen hours ahead are also examined:

2HA

$$FL_d = f(FL_u(t-2), FL_u(t-3), ..., FL_u(t-48)) \quad (8)$$

3HA

$$FL_d = f(FL_u(t-3), FL_u(t-4), ..., FL_u(t-48)) \quad (8)$$

6HA

$$FL_d = f(FL_u(t-6), FL_u(t-7), ..., FL_u(t-48)) \quad (8)$$

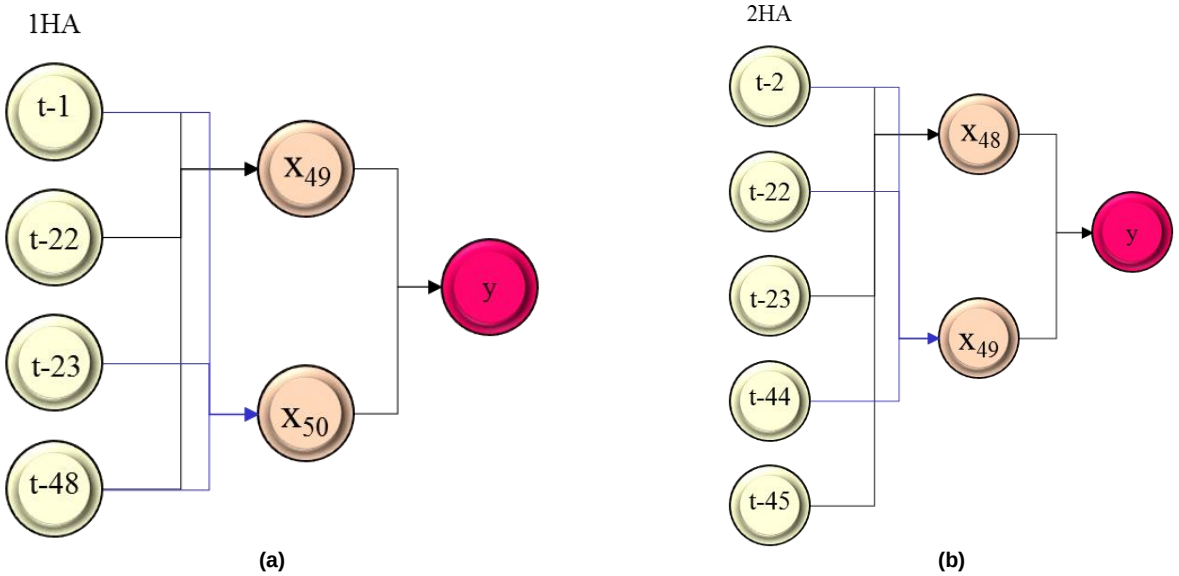
12HA

$$FL_d = f(FL_u(t-12), FL_u(t-13), ..., FL_u(t-48)) \quad (8)$$

18HA

$$FL_d = f(FL_u(t-18), FL_u(t-19), ..., FL_u(t-48)) \quad (8)$$

Therefore, the number of inputs in 2HA, 3HA, 6HA, 12HA, and 18HA equals 47, 46, 43, 37, and 31, respectively. It should be noted that limiting xHA forecasting (x = 1, 2, 3, 6, 12, 18) to 18 is because increasing this amount reduces the model's reliability, and the model could not accurately predict test data. The structure of developed GSGMDH-based models for 1HA, 2HA, 3HA, 6HA, 12HA, and 18HA is provided in Fig. 11. Based on this Fig., the number of input variables in all structures except 2HA has only four inputs. The inputs of the 2HA are equal to five. The variation in the number of inputs across different models can be attributed to the GSGMDH method's inherent capability to autonomously identify the most crucial inputs for the problem, guided by the AIC index. This criterion adeptly balances model simplicity with precision, ensuring the selection of an optimal model configuration. Besides, there are two newly generated neurons in all structures. Therefore, the developed GSGMDH model is simple and can easily be applied to practical tasks.



(a) **(b)**

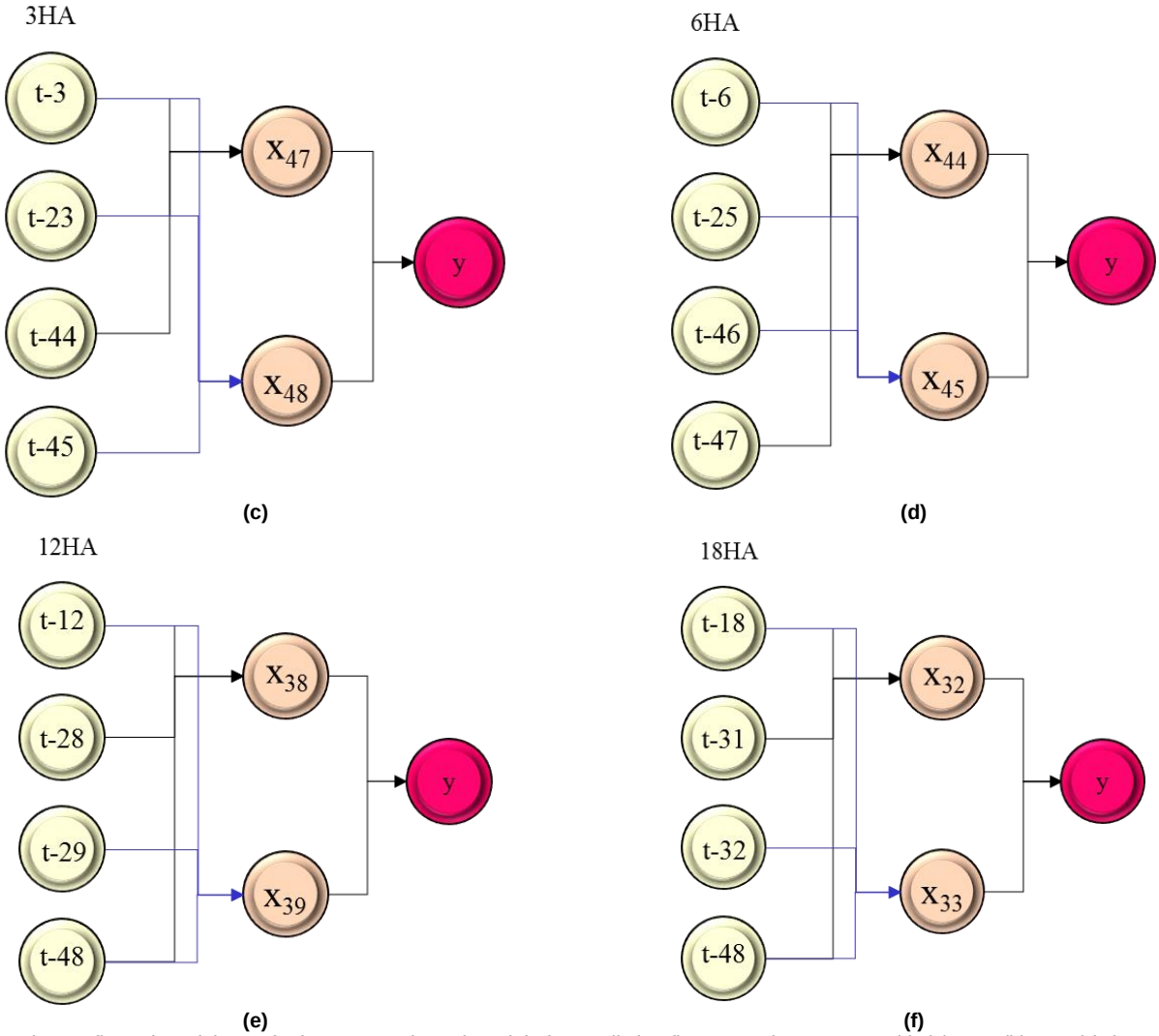


Fig. 11. The configuration of the crafted GSGMDH-based models for predicting flow rates downstream with (a) one, (b) two, (c) three, (d) six, (e) twelve, and (f) eighteen hours ahead.

Fig. 12 shows the statistical indices (i.e., R, MAE, RMSRE, NRMSE) for developed GSGMDH-based models with one, two, three, six, twelve, and eighteen hours ahead. The range of correlation coefficient index for the different models is [0.9, 0.92]. The lowest one is related to the 18HA, while for this index for 1HA, 2HA, 3HA, and 6HA, the R is almost equal to 0.92. Indeed, the change of hours ahead forecasting from one to 18 has not had a noteworthy effect on the outcomes, so the difference between the lowest and highest values of this index is about 0.02, and the the measure of this index across all models is more than 0.9, which is an acceptable value for the correlation coefficient.

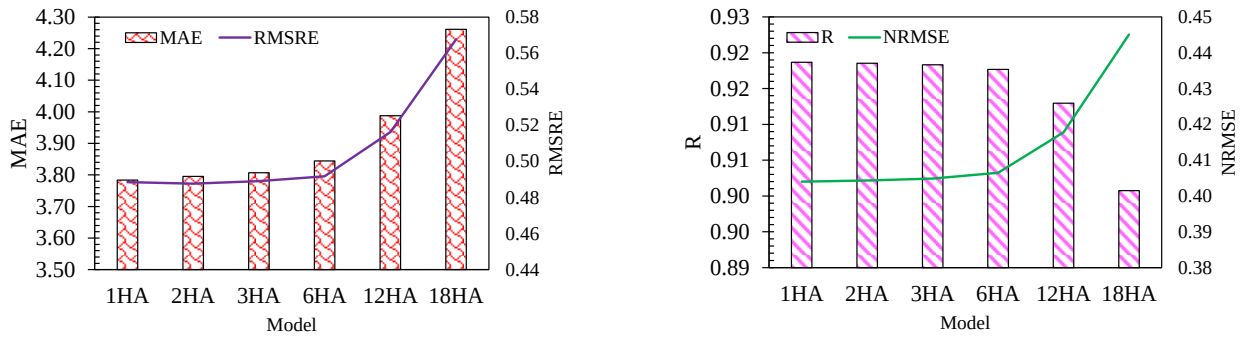


Fig. 12. Statistical indices for developed GSGMDH-based models.

The range of MAE for different developed GSGMDH-based models is [3.78, 4.26]. The increase in hours ahead forecasting is directly related to the value of this index, so the smallest MAE is linked to 1HA, while the largest corresponds to 18HA. This trend is also repeated in the NRMSE and RMSRE. The ratio of the maximum of NRMSE, MAE, and RMSRE to the minimum recorded value of each equals .1.1, 1.13, and 1.16, respectively. Due to the low value of NRMSE and RMSRE, which are very close to zero, and the correlation coefficient that is more than 0.9, it could be concluded that the developed GSGMDH-based model performs well in early-warning flood forecasting.

In addition to the provided indices in the previous Fig., the performance of those is compared with the actual values of the downstream station in terms of Boxplot, as represented in Fig. 13.

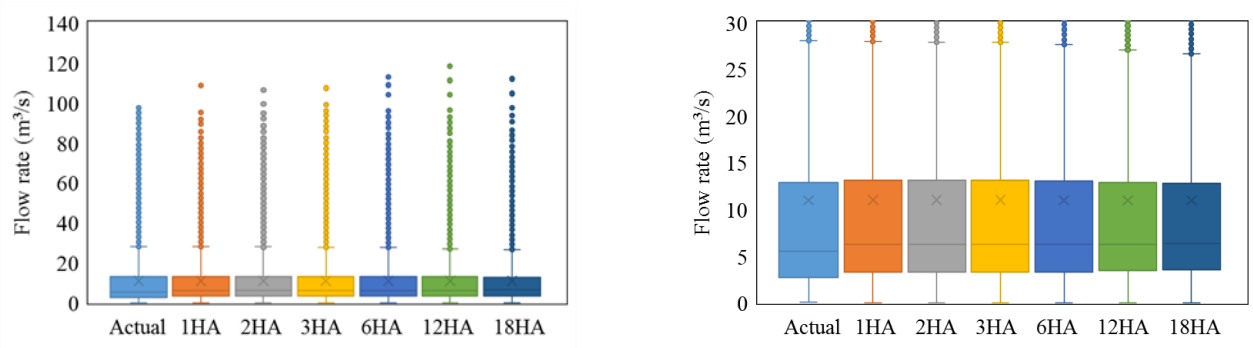


Fig. 13. Box plots for actual estimated flow rates using GSGMDH.

Due to this Fig., the distribution of the actual flow rate and forecasted ones by GSGMDH-based models are very close together. For example, the third quartile (Q3) of the actual values is 12.89 while it is 13.16, 13.12, 13.13, 13.05, 12.88, and 12.8 for 1HA, 2HA, 3HA, 6HA, 12Ha, and 18HA, respectively. The maximum difference between the Q3 of the actual values and GSGMDH-based ones is 0.27, almost 2% of the Q3 of the actual values. Moreover, the maximum difference between the maximum of the actual values and GSGMDH-based ones is 1.44 (=28.11-26.67), which is 5% of the maximum of the actual values. Besides, this Fig. shows that the peak flow rates of the actual ones are well estimated by the developed GSGMDH-based models so that even when these models do not accurately predict the actual amount, they have an overestimated performance with little difference from the actual values. Consequently, it can be concluded that the results of this Fig. confirm the the effectiveness of the models formulated in this investigation.

Fig. 14 compares the performance results for the reliability and uncertainty analysis. The reliability and uncertainty analysis details are provided in literature (Azimi, Bonakdari, Ebtehaj, 2017; Saberi-Movahed, Najafzadeh, Mehrpooya, 2020; and Azimi et al. (2017). The highest value of the reliability and lowest of the uncertainty analysis indicate the higher performance of a model. Due to this Fig., the reliability of the 1HA and 2HA is the highest, while the uncertainty of 1HA, 2HA, 3HA, and 6HA is the lowest. The weakest performance in reliability and uncertainty analysis is related to the 18HA, while the best is related to the 1HA. According to the provided values of the reliability and uncertainty analysis (Azimi, Bonakdari, Ebtehaj, 2017; Saberi-Movahed et al. 2020), the performance of the developed GSGMDH-based models is acceptable.

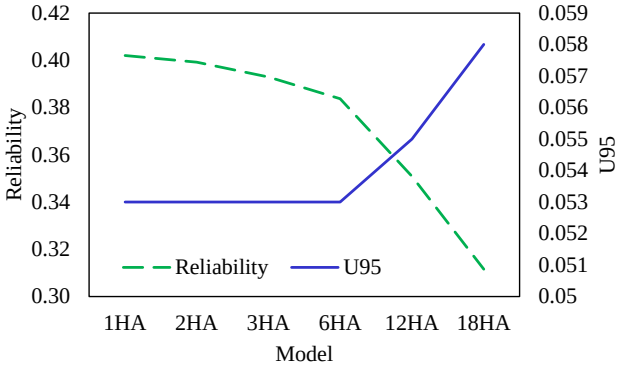


Fig. 14. Comparisons of the performance results for the reliability and uncertainty analysis.

3.2. Advantages, limitations, and future improvements

The application of the AI model in the present study proved highly effective for early-warning flood forecasting, particularly when tested on the Saint-Charles River, a significant river system in Quebec, Canada. A primary benefit of the AI models that were introduced is that they automatically determine the number of problem inputs to form a connection between the selected most essential input variables and the target variable, resulting in a simple model with large input variables defined by a user. The benefits of the suggested AI approach in this study, when contrasted with established AI-based methods, are twofold: (1) training time and (2) input variables selection.

In the current study, the developed AI model utilized least square algorithms to find the coefficients of the second and/or third-order polynomials. Nonetheless, integrating the suggested AI methodology could also be pursued in a subsequent investigation using the genetic algorithm, particle swarm optimization (PSO), gravitational search algorithm (GSA), and other evolutionary-based algorithms to optimize these coefficients. Hence, a separate investigation comparing the least square algorithm and various evolutionary algorithms in optimizing the coefficients of the polynomials is warranted. Another limitation of this study is that it uses only one upstream station as the model input for early-warning flood forecasting at the downstream station. Due to the significant difference in flood discharge in the two stations (Hurons and Saint-Charles), it seems that the input of the downstream station, in addition to the flow rate in the upstream station, is also dependent on other flows, including the flow of the other upstream stations or precipitations at these stations.

It is vital to consider the issues and challenges that prevent making progress in applying AI in flood forecasting. Some of the challenges that AI scholars may face in undertaking problems relating to visualizing, analyzing, and predicting the flood are mentioned in the following:

(1) Data challenges

(i) Missing data: About 40,000 data (15-minute recorded flow rates) were missed in the current study. Most are related to January to March, and some are in December and April. In time series-based modeling, samples must be without any missing to consider the correlation of each sample with previous ones. The input variables from one to 48 delays were generated to overcome this challenge in the current study. After that, rows with a minimum of one missing were removed. It's important to acknowledge that a significant quantity of missing samples could be impacted by the modeling results, so the generalizability of this model in forecasting the flow rate at the future time in days with most of the missing samples may be reduced.

(ii) Spatio-temporal data: Data for the current study was collected from 2008 to 2020 at only two hydrometric stations. It is highly recommended to extend data from previous years of 2008 to train an AI-based model with more extreme values of the flow rates and make its generalizability stronger. Besides, employing only one upstream station to estimate the flow rate at the downstream station cannot be considered the effect of other upstream stations. It may affect the model's performance in future times.

(iii) Data collection: The data collected in the current study were gained from hydrometric stations in Quebec, Canada, but there was missing data. To overcome the limitation of these data, it is recommended to employ satellite-based samples that are recorded in the desired time scale.

(2) Artificial Intelligence challenges

(i) Modeling in the existence of missing data: In the current, different lags of the flow rate as upstream stations were considered independent

inputs to estimate the flow rate at the downstream station. Due to that, the collected data may be missing, and the defined inputs and corresponding output may be missing data that needs to be removed from the corresponding row. Existing a large number of missing data may lead to lower generalizability of the developed model in the time scale with lower training samples.
(ii) Identifying outliers: The outliers that have a remarkable deviation from the normal group or majority of samples can be attributed to extreme events and/or imperfect collection methods. The outliers of the collected data are related to the extreme events (i.e., flooding)

4. Conclusions

Undoubtedly, one of the most devastating natural disasters is floods, experienced by different parts of the world every year. Due to the significant increase in the frequency of heavy rains, continuously enhancing assets and population concentrations in flood-prone zones, and changes in upstream land use, flood damage has increased exponentially in recent decades. According to many floods in Quebec, such as the 2019 flooding that resulted in flooding of more than 6000 residents and more than 13000 disaster victims, early-warning flood forecasting is a vital task in this area. Therefore, the flow rate relationship at these stations was checked using hourly data at two stations, Hurons and Saint-Charles, upstream of Quebec City. The primary discoveries of the present investigation include:
- Heavy rain and snowmelt are the leading causes of flooding in the area, with no clear correlation between peak flow rates at the mentioned stations.
- A novel AI-driven model was devised to forecast the flow rate at downstream stations (e.g., Saint-Charles) based on data from upstream stations (e.g., Hurons).
- The model, designed for early-warning flood forecasting, considered 48 different inputs, including lags of 1 to 48 hours of the upstream station's flow rate.
- Various models were developed for forecasting 1, 2, 3, 6, 12, and 18 hours ahead (1HA, 2HA, 3HA, 6HA, 12HA, and 18HA) at downstream stations.
- The modeling results showed that all developed AI models had a correlation coefficient of over 0.9, suggesting the model's potential as a tool for early-warning flood forecasting in Quebec.

Author Contributions

Isa Ebtehaj: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, visualization, supervision
Hossien Bonakdari: Conceptualization, methodology, validation, formal analysis, resources, data curation, writing—review and editing, project administration
Baram Gharabaghi: Writing—review and editing.

Conflict of Interest

Not applicable.

Acknowledgements

The authors would like to thank the “Ministère de l’Environnement et de la Lutte contre les changements climatiques, de la Faune et des Parcs” of Quebec, Canada.

Data Availability Statement

A third party provided all data used during the study. Direct requests for these materials may be made to the provider, as indicated in the Acknowledgments.

References

Azimi, H., Bonakdari, H., Ebtehaj, I. (2017) ‘A highly efficient gene expression programming model for predicting the discharge coefficient in a side weir along a trapezoidal canal’, *Irrigation and Drainage*, 66 (4), pp. 655-666. doi: https://doi.org/10.1002/ird.2127

Azimi, H. *et al.* (2018) ‘Evolutionary design of generalized group method of data handling-type neural network for estimating the hydraulic jump roller length, *Acta Mechanica*’, 229 (3), pp. 1197-1214. doi: https://doi.org/10.1007/s00707-017-2043-9

Balica, S.F. *et al.* (2013) ‘Parametric and physically based modelling techniques for flood risk and vulnerability assessment: A comparison’, *Environmental Modelling and Software*, 41, pp. 84–92. doi: https://doi.org/10.1016/j.envsoft.2012.11.002.

Ebtehaj, I., and Bonakdari, H. (2016) ‘A support vector regression-firefly algorithm-based model for limiting velocity prediction in sewer pipes’, *Water Science and Technology*, 73 (9), pp. 2244-2250. doi: https://doi.org/10.2166/wst.2016.064

Ebtehaj, I., and Bonakdari, H. (2023), ‘A comprehensive comparison of the fifth and sixth phases of the coupled model intercomparison project based on the Canadian earth system models in spatio-temporal variability of long-term flood susceptibility using remote sensing and flood frequency analysis’, *Journal of Hydrology*, 617, p. 128851. doi: https://doi.org/10.1016/j.jhydrol.2022.128851

Ebtehaj, I., Bonakdari, H., and Zaji, A. H. (2018), ‘A new hybrid decision tree method based on two artificial neural networks for predicting sediment transport in clean pipes, *Alexandria Engineering Journal*, 57(3), pp. 1783-1795. doi: https://doi.org/10.1016/j.aej.2017.05.021

Ebtehaj, I. *et al.* (2023a), ‘Short-Term Precipitation Forecasting Based on the Improved Extreme Learning Machine Technique’, *Environmental Sciences Proceedings*, 25(1), pp. 1-7. doi: https://doi.org/10.3390/ECWS-7-14237

Ebtehaj, I. et al. (2023b), ‘Multi-depth daily soil temperature modeling: meteorological variables or time series?’, *Theoretical and Applied Climatology*, 151(3-4), pp. 989-1012. doi: https://doi.org/10.1007/s00704-022-04314-y

Ebtehaj, I. *et al.* (2020), ‘Evaluation of preprocessing techniques for improving the accuracy of stochastic rainfall forecast models,’ *International Journal of Environmental Science and Technology*, 17(1), pp. 505-524. doi: https://doi.org/10.1007/s13762-019-02361-z

Gao, J., Christensen, P., and Li, W. (2017), ‘Application of the WEAP model in strategic environmental assessment: Experiences from a case study in an arid/semi-arid area in China’, *Journal of Environmental Management*, 198, pp. 363-371. doi: https://doi.org/10.1016/j.jenvman.2017.04.068

Gizaw, M.S.; and Gan, T.Y. (2016) ‘Regional flood frequency analysis using support vector regression under historical and future climate’, *Journal of Hydrology*, 538, pp. 387–398. doi: https://doi.org/10.1016/j.jhydrol.2016.04.041

Khozani, Z. S., Bonakdari, H., and Ebtehaj, I. (2017), ‘An analysis of shear stress distribution in circular channels with sediment deposition based on Gene Expression Programming’, *International Journal of Sediment Research*, 32(4), pp. 575-584. doi: https://doi.org/10.1016/j.ijsrc.2017.04.004

Kim, B. *et al.* (2015) ‘Urban flood modeling with porous shallow-water equations: A case study of model errors in the presence of anisotropic porosity’, *Journal of Hydrolgy*, 523, pp. 680–692. doi: https://doi.org/10.1016/j.jhydrol.2015.01.059

Kollet, S.J., and Maxwell, R.M. (2006), ‘Integrated surface–groundwater flow modeling: A free-surface overland flow boundary condition in a parallel groundwater flow model’, *Advances in Water Resources*, 29, 945–958. doi: https://doi.org/10.1016/j.advwatres.2005.08.006

Letessier, C. *et al*. (2023) ‘Enhancing flood prediction accuracy through integration of meteorological parameters in river flow observations: A case study Ottawa River’, *Hydrology*, 10(8), pp. 1-23. doi: https://doi.org/10.3390/hydrology10080164

Li, D. *et al.* (2018) ‘Development and integration of sub-daily flood modelling capability within the SWAT model and a comparison with XAJ model’ *Water*, 10(9), pp. 1263. doi: https://doi.org/10.3390/w10091263

López, P. L. *et al.* (2020) ‘Evaluation of global water resources reanalysis data for estimating flood events in the Brahmaputra River basin’, *Water Resources Management*, 34, pp. 2201-2220. doi: https://doi.org/10.1007/s11269-020-02546-z

Lotfi, K., *et al.* (2020) ‘A novel stochastic wastewater quality modeling based on fuzzy techniques’, *Journal of Environmental Health Science and Engineering*, 18, pp. 1099-1120. doi: https://doi.org/1099-1120. 10.1007/s40201-020-00530-8

Lotfi, K. *et al.* (2019) ‘Predicting wastewater treatment plant quality parameters using a novel hybrid linear-nonlinear methodology’, *Journal of Environmental Management*, 240, pp. 463-474. doi: https://doi.org/10.1016/j.jenvman.2019.03.137

McGrath, H., Stefanakis E., and Nastev., M. (2014), 'Development of a data warehouse for riverine and coastal flood risk management', *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences,* 40 (2), pp. 41-48. doi: https://doi.org/10.5194/isprsarchives-XL-2-41-2014

Mekanik, F. *et al.* (2013) 'Multiple regression and artificial neural network for long-term rainfall forecasting using large scale climate modes', *Journal of Hydrology*, 503, pp. 11–21. doi: http://doi.org/10.1016/j.jhydrol.2013.08.035

Mosavi, A. and Edalatifar, M., 2019. A hybrid neuro-fuzzy algorithm for prediction of reference evapotranspiration. In *Recent Advances in Technology Research and Education: Proceedings of the 17th International Conference on Global Research and Education Inter-Academia–2018 17* (pp. 235-243). Springer International Publishing.. doi: https://doi.org/10.1007/978-3-319-99834-3_31

Puttinaovarat, S. and Horkaew, P., 2020. Flood forecasting system based on integrated big and crowdsource data by using machine learning techniques. IEEE Access, 8, pp.5885-5905. doi: https://doi.org/10.1109/ACCESS.2019.2963819

Ramly, S. *et al*. (2020) 'Flood estimation for SMART control operation using integrated radar rainfall Input with the HEC-HMS model', *Water Resources Management*, 34(10), pp. 3113-3127. doi: https://doi.org/10.1007/s11269-020-02595-4

Saberi-Movahed, F., Najafzadeh, M., and Mehrpooya, A. (2020) 'Receiving more accurate predictions for longitudinal dispersion coefficients in water pipelines: training group method of data handling using extreme learning machine conceptions', *Water Resources Management*, 34(2), pp. 529-561. doi: https://doi.org/10.1007/s11269-019-02463-w

Safari, M. J. S. *et al.* (2019) 'Sediment transport modeling in rigid boundary open channels using generalize structure of group method of data handling', *Journal of Hydrology*, 577, p. 123951. doi: https://doi.org/10.1016/j.jhydrol.2019.123951

Sahraei, S., Asadzadeh, M., and Unduche, F. (2020) 'Signature-based multi-modelling and multi-objective calibration of hydrologic models: Application in flood forecasting for Canadian Prairies', *Journal of Hydrology*, 588, p. 125095. doi: https://doi.org/10.1016/j.jhydrol.2020.125095

Shrubsole, D. *et al.* (1993) 'The history of flood damages in Ontario', *Canadian Water Resources Journal*, 18, pp. 133–143. doi: https://doi.org/10.1016/j.jhydrol.2020.12509510.4296/cwrj1802133

Sihag, P. *et al.* (2019) 'Modeling unsaturated hydraulic conductivity by hybrid soft computing techniques', *Soft Computing*, 23, pp. 12897-12910. doi: https://doi.org/10.1007/s00500-019-03847-1

Soltani, K., *et al.* (2021) 'Mapping the spatial and temporal variability of flood susceptibility using remotely sensed normalized difference vegetation index and the forecasted changes in the future', *Science of The Total Environment*, 770, p. 145288. doi: https://doi.org/10.1016/j.scitotenv.2021.145288

Taherei Ghazvinei, P. *et al.* (2018) 'Sugarcane growth prediction based on meteorological parameters using extreme learning machine and artificial neural network', *Engineering Applications of Computational Fluid Mechanics*, 12, 738–749. doi: https://doi.org/10.1080/19942060.2018.1526119

Walton, R. *et al.* (2019) 'Estimating 2-year flood flows using the generalized structure of the Group Method of Data Handling', *Journal of Hydrology*, 575, pp. 671-689. doi: https://doi.org/10.1016/j.jhydrol.2019.05.068

Zahmatkesh, Z. *et al.* (2019) 'An overview of river flood forecasting procedures in Canadian watersheds', *Canadian Water Resources Journal/Revue canadienne des ressources hydriques*, 44(3), pp. 213-229. doi: https://doi.org/10.1080/07011784.2019.1601598

Zaji, A. H., Bonakdari, H., and Gharabaghi, B. (2018) 'Applying upstream satellite signals and a 2-D error minimization algorithm to advance early warning and management of flood water levels and river discharge', *IEEE Transactions on Geoscience and Remote Sensing*, 57(2), pp. 902-910. doi: https://doi.org/10.1109/TGRS.2018.2862640

Zeynoddin, M. *et al*. (2019) 'A reliable linear stochastic daily soil temperature forecast model', *Soil and Tillage Research*, 189, pp. 73-87. doi: https://doi.org/10.1016/j.still.2018.12.023