

## NRC Publications Archive Archives des publications du CNRC

### Estimation of instantaneous peak flows in Canadian rivers: an evaluation of conventional, nonlinear regression, and machine learning methods

Khaliq, Muhammad Naveed

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. / La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.

For the publisher's version, please access the DOI link below. / Pour consulter la version de l'éditeur, utilisez le lien DOI ci-dessous.

#### **Publisher's version / Version de l'éditeur:**

<https://doi.org/10.2166/wst.2024.096>

*Water Science & Technology*, 89, 9, pp. 2225-2239, 2024-04-17

#### **NRC Publications Archive Record / Notice des Archives des publications du CNRC :**

<https://nrc-publications.canada.ca/eng/view/object/?id=28306c5e-55a3-4c82-9a6c-c54f4b4d6484>

<https://publications-cnrc.canada.ca/fra/voir/objet/?id=28306c5e-55a3-4c82-9a6c-c54f4b4d6484>

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at

<https://nrc-publications.canada.ca/eng/copyright>

READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site

<https://publications-cnrc.canada.ca/fra/droits>

LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

**Questions?** Contact the NRC Publications Archive team at

PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

**Vous avez des questions?** Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

## Estimation of instantaneous peak flows in Canadian rivers: an evaluation of conventional, nonlinear regression, and machine learning methods

Muhammad Naveed Khaliq 

National Research Council Canada, Ocean Coastal and River Engineering Research Centre, Ottawa, ON K1A 0R6, Canada  
E-mail: muhammad.khaliq@nrc-cnrc.gc.ca

 MNK, 0000-0003-0451-7480

### ABSTRACT

Instantaneous peak flows (IPFs) are often required to derive design values for sizing various hydraulic structures, such as culverts, bridges, and small dams/levees, in addition to informing several water resources management-related activities. Compared to mean daily flows (MDFs), which represent averaged flows over a period of 24 h, information on IPFs is often missing or unavailable in instrumental records. In this study, conventional methods for estimating IPFs from MDFs are evaluated and new methods based on the nonlinear regression framework and machine learning architectures are proposed and evaluated using streamflow records from all Canadian hydrometric stations with natural and regulated flow regimes. Based on a robust model selection criterion, it was found that multiple methods are suitable for estimating IPFs from MDFs, which precludes the idea of a single universal method. The performance of machine learning-based methods was also found reasonable compared to conventional and regression-based methods. To build on the strengths of individual methods, the fusion modeling concept from the machine learning area was invoked to synthesize outputs of multiple methods. The study findings are expected to be useful to the climate change adaptation community, which currently heavily relies on MDFs simulated by hydrologic models.

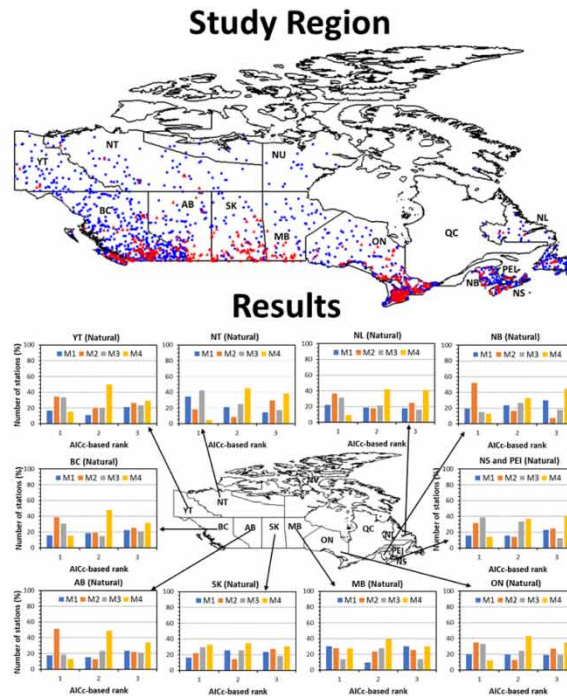
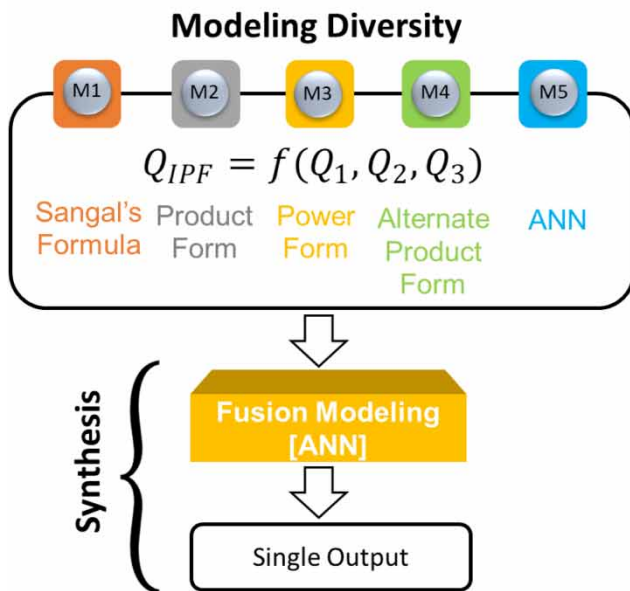
**Key words:** design flow magnitudes, fusion modeling, instantaneous peak flows, machine learning, missing values, multiple regression

### HIGHLIGHTS

- New methods for estimating instantaneous peak flows from mean daily flows.
- Data completion by filling in missing values of instantaneous peak flows.
- Reliable estimation of design flood magnitudes.
- Machine learning inspired fusion modeling to synthesize outputs of multiple methods.
- Fulfilled a critical need of the climate change impact analysis and adaptation community.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence (CC BY-NC-ND 4.0), which permits copying and redistribution for non-commercial purposes with no derivatives, provided the original work is properly cited (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

## GRAPHICAL ABSTRACT



## 1. INTRODUCTION

Instantaneous peak flows (IPFs) are often required to derive design values for sizing various hydraulic structures, such as culverts, bridges, small dams, and levees, in addition to informing hydraulic procedures of floodplain delineations, erosion risk management strategies, and water resources management-related activities. Statistical frequency analysis approaches, which are commonly used for deriving targeted design flow magnitudes, rely considerably on long sequences of observed IPFs. However, compared to IPFs, information on mean daily flows (MDFs) is more frequently available at hydrometric stations, both in Canada and other regions of the world. For a given year or season, IPF is the maximum recorded flow at the crest of the hydrograph at the point of interest within a river and is generally associated with a fine temporal scale such as 5- or 10-min interval. In comparison, MDF represents an averaged aggregated quantity corresponding to a continuous period of 24 h (or a calendar day) and is generally expected to be smaller than the IPF due to the averaging involved. Consequently, design flow values derived from extreme observations of MDFs need to be adjusted upward based on established techniques or using region- or watershed-specific empirical experience.

In cases where long sequences of IPFs are available, individual missing values and continuous gaps of missing values are often encountered in observational records. To overcome such situations, investigators have designed different approaches to estimate IPFs from MDFs (e.g., Sangal & Kallio 1977; Sangal 1981; Chen *et al.* 2017). This estimation using a reliable approach has become important recently for evaluating climate resilience of various flood control and mitigation structures in the face of changing climatic conditions. According to the International Panel on Climate Change (IPCC) (IPCC 2013), it is very likely that the frequency and magnitude of floods in certain regions of the world, including Canada, will increase due to the anticipated climate change. In order to evaluate such impacts of climate change and outline/plan adaptation strategies, climate model outputs of temperature, precipitation, and other relevant variables are integrated with hydrological models to simulate streamflow sequences for a selected reference period and many future time horizons. These simulated sequences from hydrological models mostly become available in the form of MDFs, and very rarely at smaller sub-hourly timescales to support estimation of IPFs. For this purpose, appropriate methods must be used to estimate IPFs from sequences of MDFs. Currently, this information is lacking in the literature. Therefore, the main objective of this paper is to develop new methods for estimating IPFs from MDFs, using data from a vast network of hundreds of hydrometric stations located across Canada, so that they can reliably be used in climate change adaptation-related studies.

The research on estimating IPFs from MDFs started decades ago. Early researchers assumed IPFs as functions of associated MDFs occurring over a short time window, which was taken as three consecutive days. In these methods, it was explicitly assumed that the IPF is a function of the maximum MDF of the sequence and select characteristics of the associated hydrograph shape, reflective of rising and falling portions of the hydrograph. Langbein (1944) developed such a graphical approach which was tested by Ellis & Gray (1966) for Canadian Prairie watersheds. Inspired by Langbein's work, Sangal (1981) introduced a quantitative approach, which was tested on observed floods from 387 natural and regulated stations from Ontario, Canada. About 70% of these stations had drainage areas less than 1,000 km<sup>2</sup> and about 50% of floods were snowmelt dominated, which is a common feature in Canada. Fill & Steiner (2003) also introduced an MDF-driven IPF estimation method, which was applied to 117 floods from 14 stations with drainage area ranging from 84 to 687 km<sup>2</sup>. Along these lines, Chen *et al.* (2017) recently introduced a slope-based method which takes into account slopes of the rising and falling portions of the hydrograph to describe the shape of the MDF hydrograph. As no parameter is required to be estimated from data, this method is much simpler than the methods of Sangal (1981) and Fill & Steiner (2003). This method was evaluated on roughly 3,800 floods observed at 144 watersheds from Iowa, with drainage area ranging from 70 to 221,700 km<sup>2</sup>. Among other methods evaluated for estimating IPFs include machine learning-based approaches (e.g., Dastorani *et al.* 2013; Jimeno-Sáez *et al.* 2017) and deterministic type conceptual and distributed hydrologic modeling (e.g., Singh 1995; Singh & Frevert 2002). The hydrologic modeling approach requires much more data and deterministic modeling expertise than the other methods mentioned above.

In this paper, Sangal's method (Sangal 1981) and three new nonlinear generalizations of the same concept as used in this method, along with an Artificial Neural Network (ANN)-based approach are evaluated using IPF and MDF data from 1,938 available natural and regulated flow measuring stations, covering nearly all provinces and northern territories of Canada. This level of country-wide study has never been attempted before. A country-wide study is useful in uncovering relative suitability of various methods in different regions as floods can originate from many different generating mechanisms across the Canadian landscape (Teufel & Sushama 2021). This combined with different geophysical and other factors can significantly impact the relationship between IPFs and MDFs. Consideration of regulated stations is useful in validating the applicability of observations made from natural flows to regulated flows, particularly during high flow conditions when the influence of regulation weakens. Performance of the studied methods is evaluated using multiple assessments and an objective model selection criterion that accounts for the structural complexity of candidate models is also introduced to aid decision-making, especially for the conventional and regression-based methods. An ANN-based fusion modeling approach is also introduced to synthesize outputs from multiple methods. The work reported in this paper is a natural extension of the pilot study conducted recently (Khaliq 2023), where the performance of the aforementioned methods was evaluated using data from natural flow measuring stations from the four Canadian Atlantic provinces, i.e., New Brunswick, Newfoundland and Labrador, Nova Scotia, and Prince Edward Island. Most of the outputs of that preliminary study are integrated in this country-wide study and for this reason, the reader may find some resemblances for the Atlantic provinces, which are also covered in this study. In addition, theoretical considerations as described in Section 2 also bear resemblance with those provided in Khaliq (2023) because most of the theoretical descriptions and principles cannot be changed.

This paper is organized as follows. Information on theoretical underpinning of the conventional and newly introduced nonlinear regression-based IPF estimation methods is described in Section 2, along with the parameter estimation and performance evaluation strategy employed in the study. Information on hydrometric data and the study area is provided in Section 3, followed by detailed results of the study in Section 4. Overall conclusions and future directions are summarized in Section 5.

## 2. METHODS

### 2.1. Methodological background

As mentioned in the introduction section, research on the estimation of IPFs started long ago perhaps due to the lack of availability of continuous water level recorders at the time or due to the desire of filling in missing values of IPFs. In the Canadian context, research on this subject was initiated by Langbein (1944) who introduced a graphical approach for estimating IPFs. Inspired by the work of Langbein (1944), Sangal (1981) introduced a quantitative method for estimating IPFs from MDFs assuming a triangular hydrograph and geometrical analogy. According to the Sangal's proposal, for a given event, the IPF

can be estimated from MDFs using the following relationship,

$$Q_{IPF} = \frac{Q_1 + Q_3}{2} + \frac{2Q_2 - Q_1 - Q_3}{(1 - 2\alpha)} \quad (1)$$

where  $Q_{IPF}$  is the IPF;  $Q_2$  is the maximum of the MDFs of a sequence of three consecutive days in the vicinity of the IPF;  $Q_1$  and  $Q_3$  are, respectively, the MDFs of the prior and succeeding days of  $Q_2$ . The term  $(1 - 2\alpha)$  was named as the base factor by Sangal (1981) and it lies between 0 and 2. Since determination of the base factor (or more specifically the parameter  $\alpha$ ) involves additional processing, a simplified relationship was also introduced in Sangal (1981) by assuming  $\alpha = 0$ . Following this assumption, the above equation can be written as:

$$Q_{IPF} = \frac{4Q_2 - Q_1 - Q_3}{2} \quad (2)$$

In the literature on this subject, the simplified Equation (2) was used more frequently than Equation (1), the original form of Sangal's method. Due to the introduced simplification, it is expected that Equation (2) may not perform as well as Equation (1) because the parameter  $\alpha$  helps in tuning the method based on observations or regional considerations. Sangal (1981) applied this procedure to 3,946 IPF events observed at 387 stations in Ontario, Canada. In many subsequent studies (e.g., Dastorani *et al.* 2013; Chen *et al.* 2017) only Equation (2) was considered, instead of evaluating the full potential of the method. It is likely to confuse the annual or seasonal maximum MDF, which is often considered for statistical frequency analyses, with the maximum MDF  $Q_2$ . These two are different quantities, however, both will coincide in the majority of the cases, especially where hydrological regimes are dominated by snowmelt spring floods. Fill & Steiner (2003) introduced an IPF estimation method by assuming IPF as a linear combination of MDFs of three consecutive days and a regression-based correction factor. The methodological framework of this method lacks theoretical robustness and therefore extreme caution should be exercised in using this method. No further consideration is given to this method in the present study. Chen *et al.* (2017) proposed a slope-based method, by assuming a triangular hydrograph. Following their proposal, the IPF as a function of MDFs is given by:

$$Q_{IPF} = Q_2 + \frac{(Q_2 - Q_1)(Q_2 - Q_3)}{2Q_2 - Q_1 - Q_3} \quad (3)$$

The various quantities involved in Equation (3) have the same meanings as in Equation (1). However, it must be noted that both Equations (2) and (3) do not require any parameter to be estimated for their application, which is an attractive feature of these methods, apart from the imposed hydrograph shape restrictions. If the shape restrictions are relaxed, then it can also be assumed that the IPF is a direct function of three MDFs occurring in the vicinity of the IPF, i.e.,  $Q_{IPF} = f(Q_1, Q_2, Q_3)$ . As a consequence of this generalization, several different linear and nonlinear forms of  $f(\cdot)$  can be considered for estimating IPFs from MDFs. As in Khaliq (2023), the following nonlinear forms are also considered in this study:

$$Q_{IPF} = aQ_1^b Q_2^c Q_3^d \quad (4)$$

$$Q_{IPF} = aQ_2^b \quad (5)$$

$$Q_{IPF} = aQ_2^b \left( \frac{Q_1 + Q_3}{2} \right)^c \quad (6)$$

Equations (4) and (6) are of multiplicative nature, while Equation (5) is of power form. The latter equation assumes that the IPF is a direct nonlinear function of the maximum MDF of the sequence of three consecutive MDFs occurring in the vicinity of the IPF. In these equations,  $Q_{IPF}$ ,  $Q_1$ ,  $Q_2$ , and  $Q_3$  are the flows as explained above; and  $a$ ,  $b$ ,  $c$ , and  $d$  are the parameters of the functional relationships which are to be estimated by iterative means or following an optimization routine subject to applicable constraints or via nonlinear least squares approach. In a watershed, IPFs are also impacted by many physical factors, such as hydrological and physiographical controls, water storage features, soil moisture state, and meteorological inputs.

These factors were not considered in this study as the focus was kept on explicit functional relationships and machine learning architectures.

Similar to [Khaliq \(2023\)](#), Sangal's method (Equation (1)) and the above mentioned nonlinear functional relationships (i.e., Equations (4)–(6)), were evaluated using data from the Canadian hydrometric network, to be discussed in Section 3. Henceforth, these methods are respectively referred to as M1, M2, M3, and M4 for presentation convenience. In addition to these four methods, a feedforward multilayer perceptron (MLP) ANN ([Maier et al. 2010](#); [Lantz 2015](#); [Chattopadhyaya et al. 2019](#)) was also considered. This method is referred to as M5. Complexity of different ANN structures and their relative performance are discussed in Section 4.2.

## 2.2. Parameter estimation and optimization of ANN structure

The parameter  $\alpha$  in Equation (1) and parameters  $a$ ,  $b$ ,  $c$ , and  $d$  of Equations (4)–(6) were estimated by minimizing root mean square error (RMSE), which is equivalent to minimizing mean squared differences between observed and estimated IPFs. The RMSE is defined as:

$$\text{RMSE} = \left[ \frac{1}{n} \sum_{i=1}^n \{Q_{IPF}^{Obs}(i) - Q_{IPF}^{Est}(i)\}^2 \right]^{1/2} \quad (7)$$

where  $Q_{IPF}^{Obs}$  is the observed and  $Q_{IPF}^{Est}$  is the estimated IPF and  $n$  is the number of sample pairs used in the optimization of parameters. Alternatively, these parameters can also be estimated following ready to use least squares software tools by linearizing Equations (4)–(6) using logarithmic transformation, which is relatively a straightforward procedure.

The MLP ANN was trained using the resilient backpropagation algorithm ([Anastasiadis et al. 2005](#)) as implemented in the 'neuralnet' package of R computing platform ([R Core Team 2023](#)) with ReLU (rectified linear unit) as the activation function, an input layer, an output layer, and one or more hidden layers. The number of hidden layers and the number of neurons in each of the hidden layers is generally determined by trial and error after evaluating the performance of different network structures ([Maier et al. 2010](#); [Wolfs & Willems 2014](#); [Lantz 2015](#)). In this study, the optimal network structure and all controlling parameters were determined using 70% of data for training and reserving 30% for testing purposes, and experimenting with multiple hidden layers containing multiple neurons.

## 2.3. Model selection criterion

Since multiple methods for estimating IPFs are considered in this study, an objective model selection criterion was necessary to be employed for identifying a better performing method from the set of all methods considered. For this purpose, Akaike Information Criterion (AIC) ([Akaike 1974](#)) was used to aid this selection. This criterion has been extensively used in the literature for selecting an optimal model from a set of competing models. Since the above methods M1–M4 involve different numbers of parameters to be estimated from observations, a corrected form of AIC was used. For regression type problems, the corrected form of AIC (denoted AICc) is given by the following relationship ([Hurvich & Tsai 1989](#); [Bonakdari & Zeynoddin 2022](#); [Burnham & Anderson 2002, 2004](#)):

$$\text{AICc} = n \ln \left( \frac{\text{SSE}}{n} \right) + 2k + \frac{2k(k+1)}{n-k-1} \quad (8)$$

In this equation, SSE is the sum of squares of errors, which is embedded in Equation (7),  $k$  is the number of estimated parameters and  $\ln$  is the natural logarithm. The second and the third term on the right penalize AIC for the number of estimated parameters and sample size. A method or a model with a smaller value of AICc is the most preferred option from a given set of modeling options. In situations where competing modeling options exhibit varying levels of structural complexities reflected in terms of controlling model parameters and small sample problems, the AICc is generally favored due to the corrections introduced ([Brewer et al. 2016](#)).

## 2.4. Performance measures

Although the above-described RMSE and AICc were, respectively, used for parameter estimation and selection of a best performing method, these measures can also be used to evaluate performance of various methods when considered in a comparative mode. In addition to these measures, mean absolute percentage error (MAPE) and the coefficient of

determination (i.e., the  $R^2$  measure) were also considered. The MAPE is given by Equation (9), while  $R^2$  is given by Equation (10):

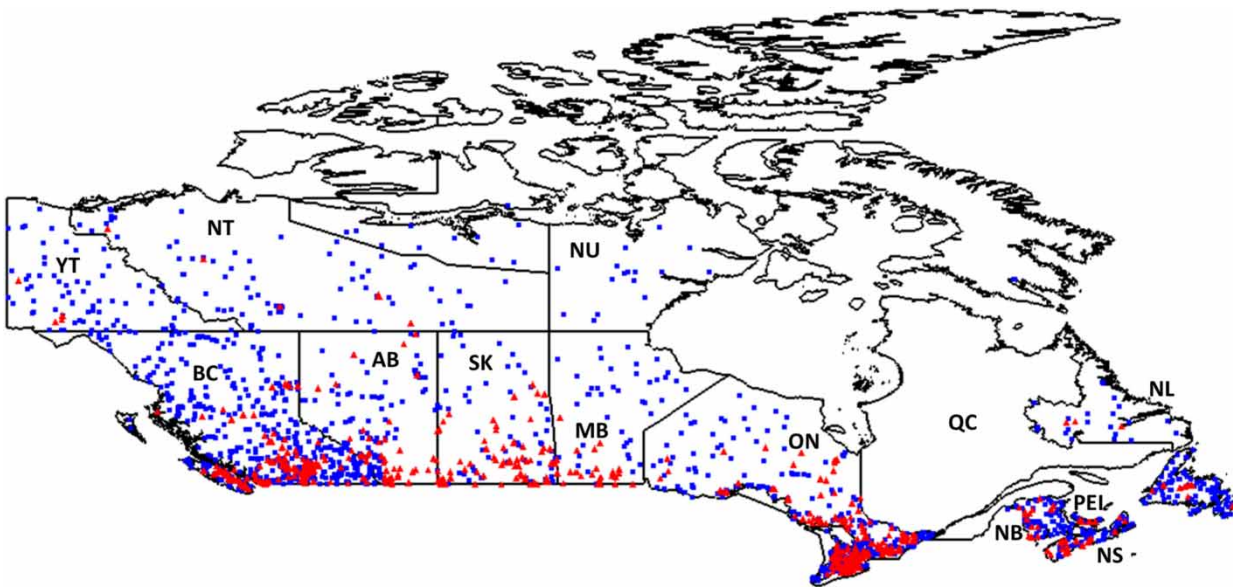
$$\text{MAPE} = \frac{1}{n} \sum_{i=1}^n \left| \frac{Q_{IPF}^{Obs}(i) - Q_{IPF}^{Est}(i)}{Q_{IPF}^{Obs}(i)} \right| \times 100 \quad (9)$$

$$R^2 = \left[ \frac{\sum_{i=1}^n [Q_{IPF}^{Obs}(i) - \overline{Q_{IPF}^{Obs}}][Q_{IPF}^{Est}(i) - \overline{Q_{IPF}^{Est}}]}{\sqrt{\sum_{i=1}^n [Q_{IPF}^{Obs}(i) - \overline{Q_{IPF}^{Obs}}]^2 \sum_{i=1}^n [Q_{IPF}^{Est}(i) - \overline{Q_{IPF}^{Est}}]^2}} \right]^2 \quad (10)$$

In Equation (10),  $\overline{Q_{IPF}^{Obs}}$  and  $\overline{Q_{IPF}^{Est}}$  are, respectively, the observed and estimated mean values. The  $R^2$  measure is included here due to its common usage in the literature, but it must be carefully interpreted with guidance from the published literature (Helsel & Hirsch 2002; Walpole *et al.* 2012). This measure can be impacted by outliers and systematic patterns in data and therefore it should be accompanied by visual comparisons as is done in this study and also demonstrated in Helsel & Hirsch (2002).

### 3. STUDY AREA AND STREAMFLOW DATA

This study is performed for all provinces and territories of Canada, except Quebec, which is not considered because of unavailability of IPF records in the hydrometric data archive at the time of the study. Henceforth, all provinces and territories are referred to as P&Ts to aid presentation and discussion. Additionally, commonly used abbreviated names of all P&Ts are used instead of full names. That is, AB for Alberta, BC for British Columbia, MB for Manitoba, NB for New Brunswick, NL for Newfoundland and Labrador, NS for Nova Scotia, NT for Northwest Territories, NU for Nunavut, ON for Ontario, PEI for Prince Edward Island, QC for Quebec, SK for Saskatchewan, and YT for Yukon Territories. A spatial plot of hydrometric stations, with both natural and regulated flow regimes, considered for this study is shown in Figure 1. All observational records until the year 2021 were considered and the starting year was not constrained to a specific year as it was not necessary for this study. Stations with less than 5 years of data and those where drainage area information was not available were excluded from the analysis. In addition, when deciding the length of available records, record years where concurrent



**Figure 1** | Location of considered hydrometric stations, with natural (blue symbols) and regulated (red symbols) flow regimes, shown on the Canadian map. The names of provinces and territories are shown using the commonly used abbreviated form, discussed in Section 3.

values of MDF and IPF were not available and those with 30% missing values were not considered. All streamflow data used in this study is sourced from the HYDAT database of Environment and Climate Change Canada (ECCC). The HYDAT database is the national archive of surface water quantity data, which is collected and maintained by ECCC through cooperative partnerships with provincial and territorial governments (ECCC 2023). Its online version along with selected streamflow statistics is also available on ECCC's web portal, [https://wateroffice.ec.gc.ca/mainmenu/historical\\_data\\_index\\_e.html](https://wateroffice.ec.gc.ca/mainmenu/historical_data_index_e.html).

P&T-wise station counts that resulted following the above criteria are given in Table 1. The station count for PEI was small and hence these stations were merged with those for NS and the same was done for the case of NU, where stations were merged with those for NT. The number of flood events or station years considered for the analysis are also shown in this table.

To quickly see the range of drainage areas of all natural and regulated watersheds, cumulative frequency curves of drainage areas were developed and those are shown in Figure 2. This figure suggests that ~53% of all natural watersheds lie between 20 and 1,000 km<sup>2</sup> and ~80% of them lie between 20 and 10,000 km<sup>2</sup>. For the same percentage points, the regulated watershed limits tend to be larger than those of natural watersheds.

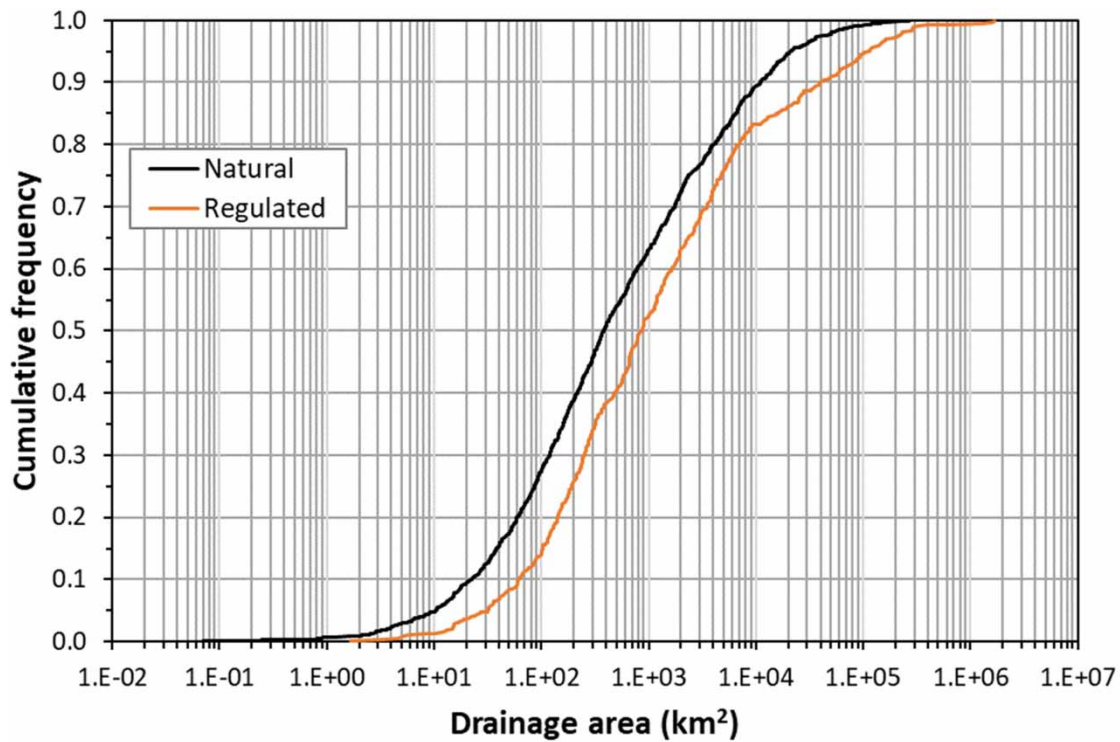
#### 4. RESULTS AND DISCUSSION

For each provincial and territorial region, two sets of analyses were conducted. One set pertains to the period of record (POR) highest IPF and associated MDFs for each of the considered stations together; the outcomes of this analysis can help develop flood envelop curves and Creager type diagrams (Creager *et al.* 1945; Neill 1986; Watt 1989) by infilling observational records of IPFs. The second set pertains to station-based analysis, conducted independently for each of the considered stations within each of the provincial and territorial region. The results of these two sets of analyses are presented and discussed below in an independent fashion.

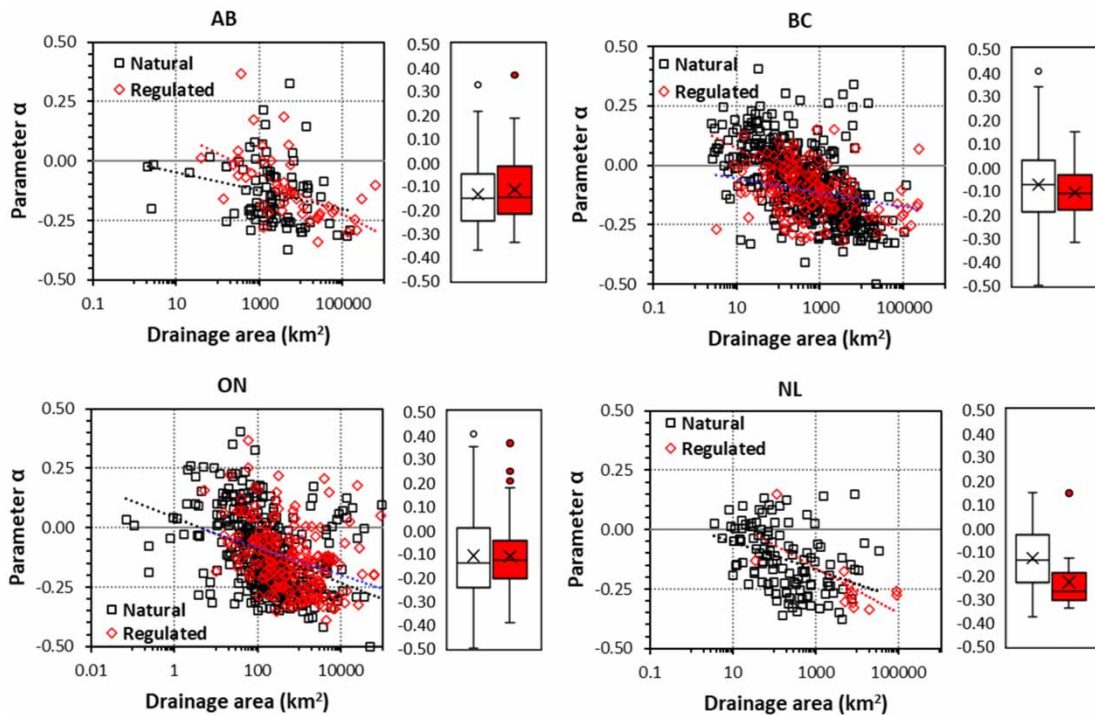
Before presenting the comparative performance of M1–M4 methods, a targeted investigation of Sangal's method (M1) is presented because this was one of the first methods which was introduced to quantitatively estimate IPFs from MDFs and therefore it deserves detailed evaluation. This method involves a single parameter  $\alpha$ , which is to be estimated from observations. This was achieved by minimizing RMSE as described in Section 2. The behavior of this parameter was investigated with respect to the watershed drainage area, as shown in Figure 3. In Figure 3 and in Figure S1 (Supplementary material), a downward trend with increasing drainage area can be seen for many P&Ts (e.g., AB, BC, ON, NL, NB, YT, and NS) for both natural and regulated watersheds, meaning a tendency toward negative values as drainage area increases. Contrary to this observation, an opposite behavior was also noticed, especially for natural watersheds (e.g., MB and SK watersheds). Given this situation, it is difficult to make a general statement for the entire country. However, the former observation holds for most P&Ts. The behavior of box plots of  $\alpha$  suggests negative median and mean values and considerable variability among estimated  $\alpha$  values. The majority of  $\alpha$  values was found negative, which was also noted in Sangal (1981). This observation leads to the conclusion that the original method of Sangal (i.e., Equation (1)) must be preferred

**Table 1** | Number of natural and regulated hydrometric stations, along with accumulated number of flood events or station years, considered in the study

Province and territory (P&T)	Natural stations		Regulated stations	
	Number of stations	Number of flood events/station years	Number of stations	Number of flood events/station years
AB	86	3,030	44	1,972
BC	435	12,593	143	4,390
MB	43	1,193	24	740
NB	67	2,024	14	517
NL	117	3,137	12	280
NS + PEI	57	1,590	26	749
NT + NU	104	1,921	6	147
ON	355	9,080	204	6,914
SK	55	1,912	70	2,941
YT	72	1,876	4	116



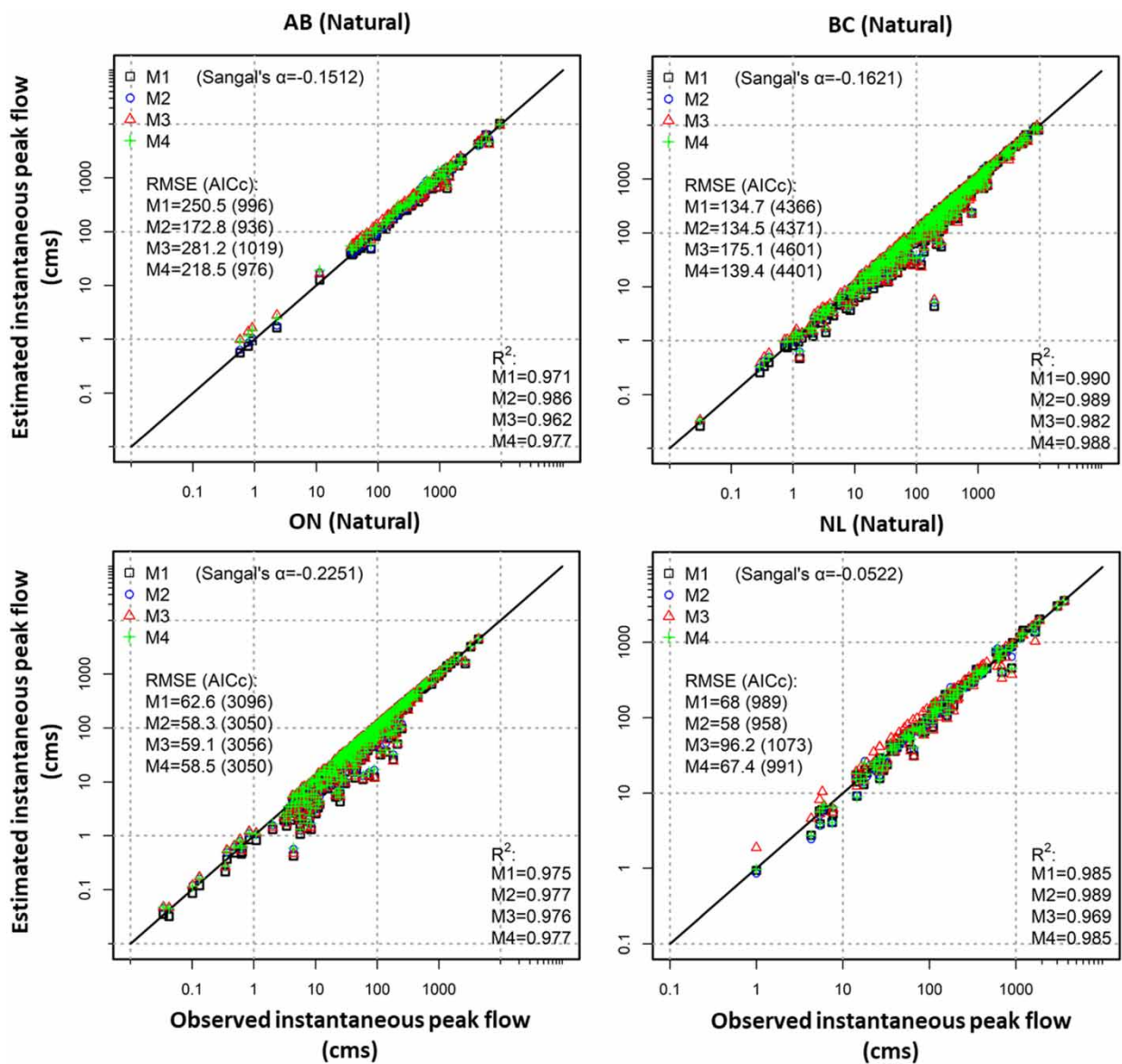
**Figure 2** | Cumulative distribution of drainage areas of both natural and regulated watersheds considered in the study.



**Figure 3** | Scatterplots of parameter  $\alpha$  of method M1 as a function of watershed drainage area for AB, BC, ON, and NL for both natural and regulated stations. Dotted lines represent corresponding trends and whisker plots represent corresponding nonparametric summary statistics with the mean values shown using a cross symbol. Whiskers correspond to minimum and maximum values, while the box represents the 25th and 75th percentile values. Median value is located inside the box. Outlying points show the values which lie outside 1.5 times the interquartile range as described in [Helsel & Hirsch \(2002\)](#).

for reliably estimating IPFs from MDFs. Although the simplified method (i.e., Equation (2)) would work in limited number of cases, it should be used cautiously. In this study, both forms of Sangal’s method were evaluated, but only the results of the original method are reported here.

For the first set of analyses, comparison of observed (POR highest) and estimated IPFs are shown in Figure 4 for M1–M4 methods. In this figure, values of RMSE,  $R^2$ , Sangal’s parameter  $\alpha$ , and AICc are also provided. It can be seen that most of the estimated and observed IPFs align well along the diagonal line, suggesting a close agreement between observed and estimated IPFs. Performance metrics also suggest quite reasonable performance of all methods as is reflected by above 0.9  $R^2$  values for all P&Ts. Based on the values of AICc, method M2 outperformed other three methods. Although four parameters are involved in this method to provide additional flexibility in fitting, their effect is taken into account via the corrected form of AIC. The performance of method M3 is quite remarkable given that IPF is considered as a nonlinear function of the maximum MDF from the sequence of three consecutive day’s MDFs. Thus, the maximum MDF alone is able to explain a



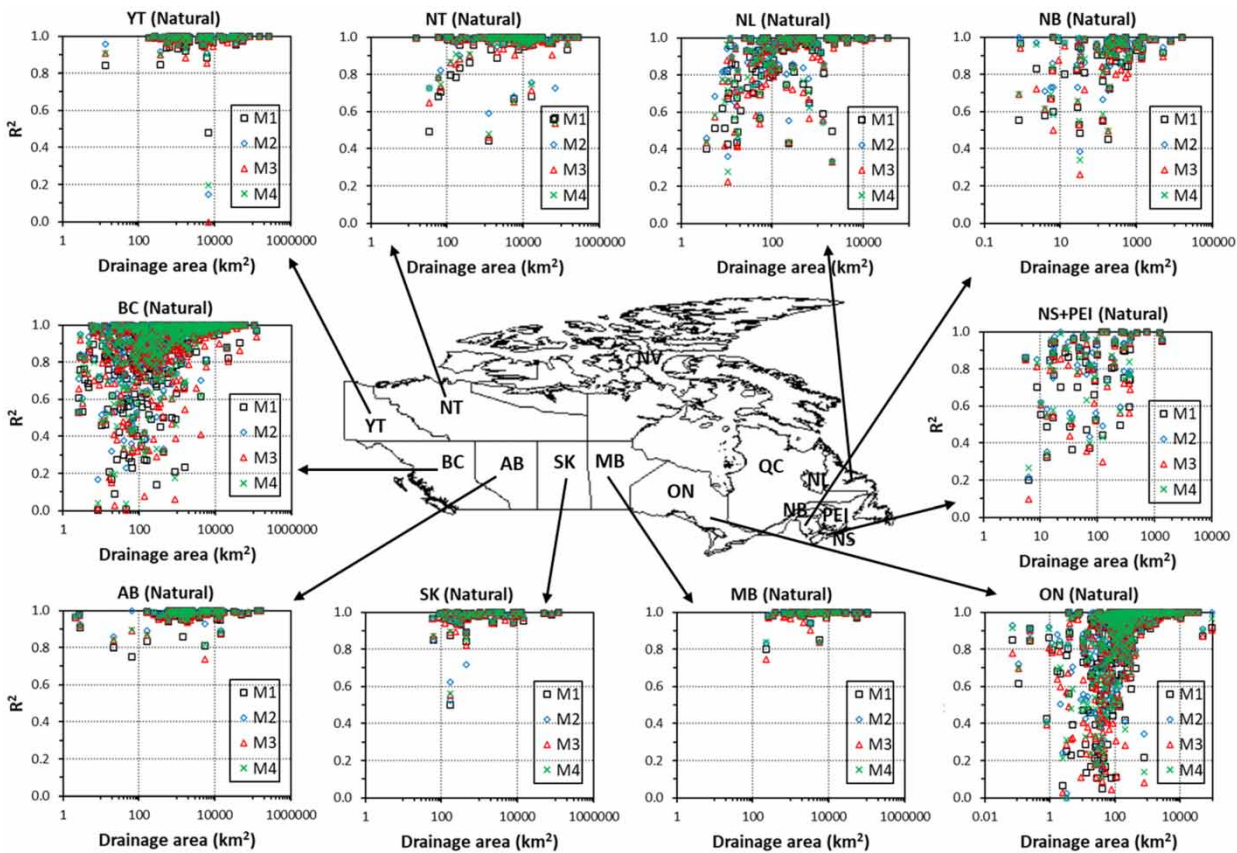
**Figure 4** | Comparison of POR highest observed and estimated IPFs for AB, BC, ON, and NL for M1–M4 methods, considering watersheds with natural flow regimes. The diagonal line represents the line of perfect match. Performance measures and model selection criterion values are embedded in all plots. Similar plots for other P&Ts are available in Figure S2, Supplementary material.

considerable portion of the observed variability in IPF values. With respect to estimated number of parameters, this method is a sort of parsimonious method compared to M2 and M3 methods. The performance of method M1, the Sangal's original method, is also reasonable, except for a few stations for certain P&Ts (e.g., NB, NS, and SK). It was also noticed that the performance of method M1 was generally better than the simplified method (i.e., Equation (2)).

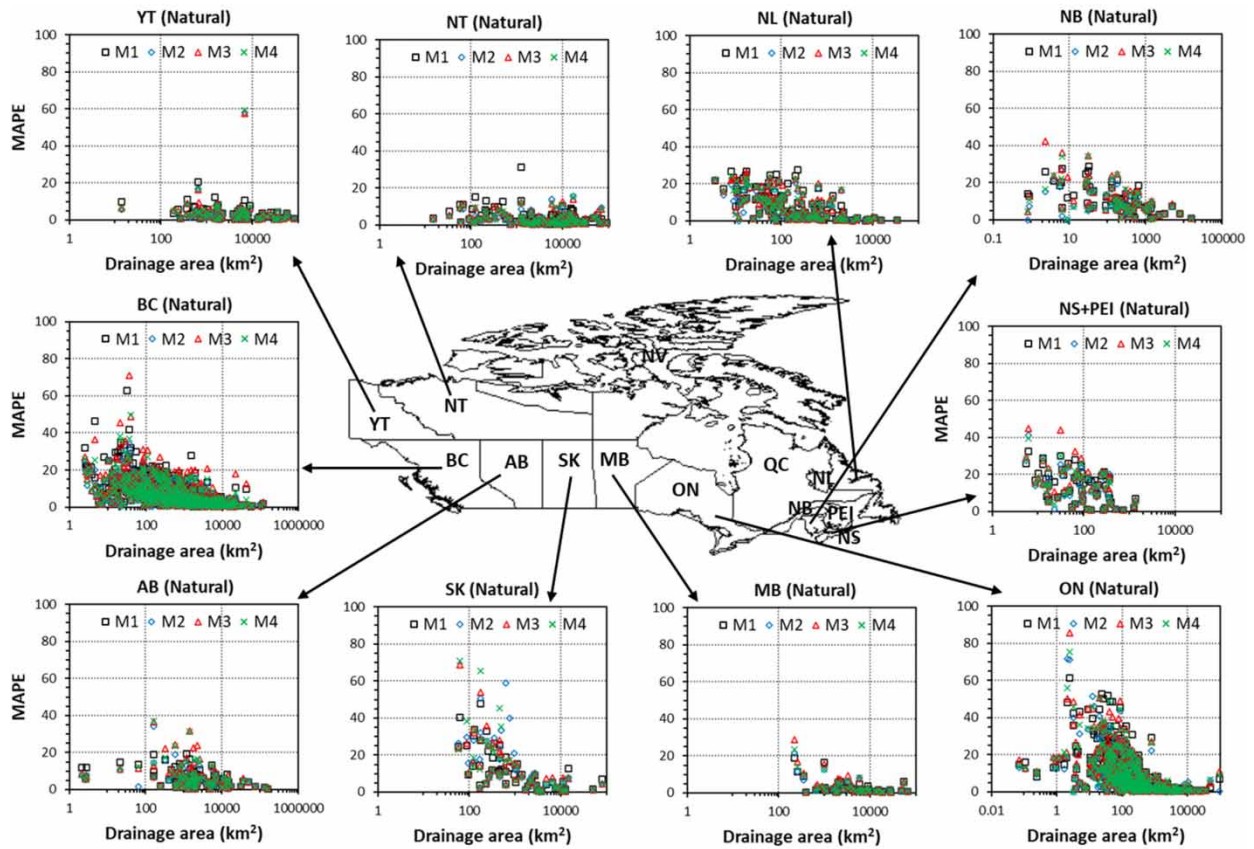
For the second set of analyses, the performance of M1–M4 methods was evaluated based on  $R^2$  and MAPE for each natural and regulated station separately within all P&Ts. The results of this evaluation based on  $R^2$ , as a function of drainage area, for natural stations are summarized in Figure 5 and those based on MAPE in Figure 6. For the case of regulated stations, similar results are provided respectively in Figures S3 and S4 (Supplementary material). Overall, scatterplots of  $R^2$  suggest values larger than 0.7 for the majority of watersheds, with generally higher degree of spread for smaller watersheds than larger ones. For AB, SK, MB, NT, and YT, the percentage of watersheds with larger than 0.8  $R^2$  is significantly higher than other P&Ts. For a small number of watersheds, especially for BC, ON, and NS, very low  $R^2$  values ( $<0.2$ ) can also be seen in Figure 5. For such cases, among other elements, data reliability or influences of local factors could be potential reasons. However, no causative analysis was conducted to investigate these issues. Side-by-side box and whisker plots of  $R^2$  for each of the four methods (see Figure S5) suggest better performance of the M2 method for most P&Ts.

Scatterplots of MAPE suggest below 20% values for most of the watersheds for all methods and P&Ts, with decreasing pattern for larger watersheds, which can clearly be seen in Figure 6 for BC, NB, SK, and ON. This means IPFs are better predicted for larger watersheds compared to smaller ones. Additional analyses based on box and whisker plots of MAPE values (Figure S6, Supplementary material) confirm these findings and suggest that the median values of MAPE are relatively smaller for method M2 than the other methods.

AICc-based rankings for the selected four methods for all P&Ts are shown in Figure 7 for stations with natural flow regimes. These rankings suggest that it is not possible to pick a single method which can perform the best for all watersheds.



**Figure 5** | Scatterplots of  $R^2$  values as a function of drainage area for methods M1–M4 for all P&Ts and watersheds, with natural flow regimes.



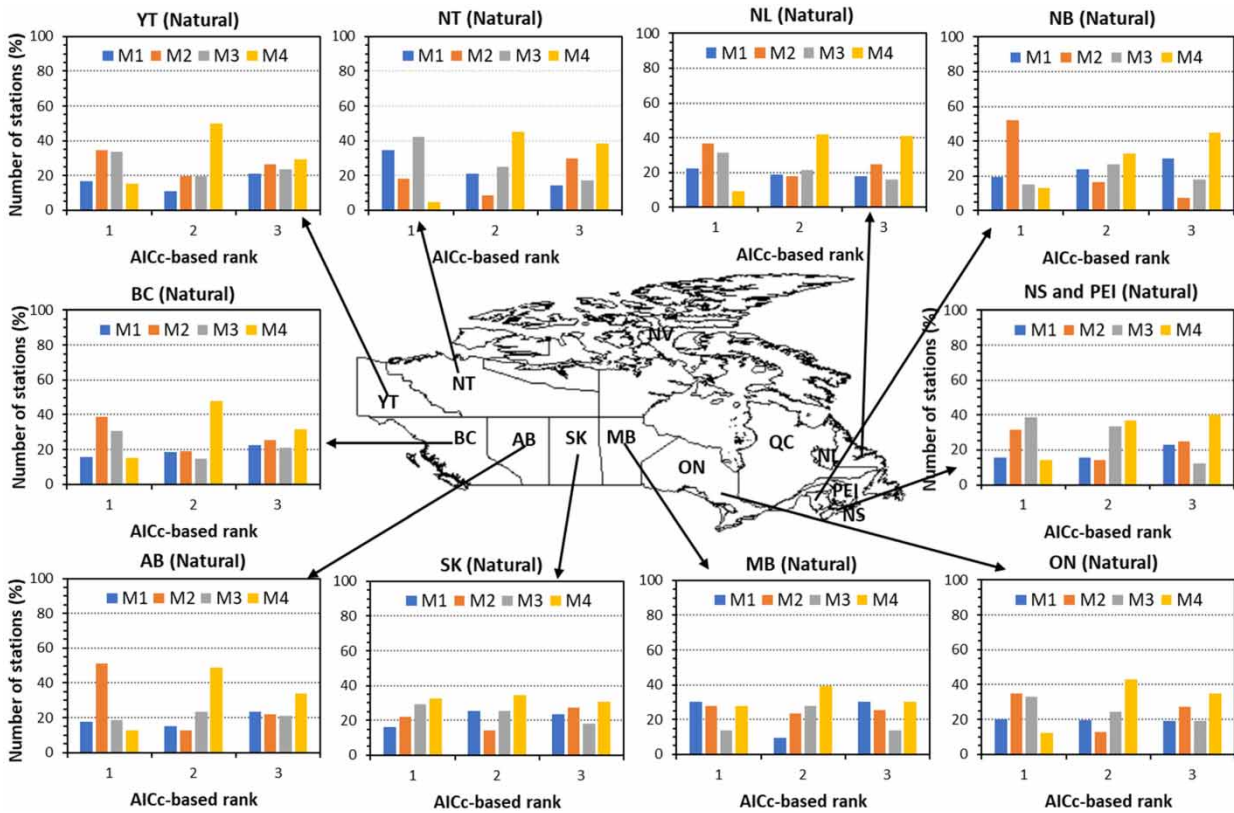
**Figure 6** | Scatterplots of MAPE values as a function of drainage area for methods M1–M4 for all P&Ts and watersheds, with natural flow regimes.

For example, M1 is the preferred method for AB, BC, NB, NL, ON, and YT, while the preferred method for NT + NU and NS + PEI is M3. Similarly, M4 is the preferred method for SK and M1 for MB. At the second and third rank, method M4 is the predominant method. Similar rankings for regulated watersheds are given in Figure S7 (Supplementary material), which generally support these findings.

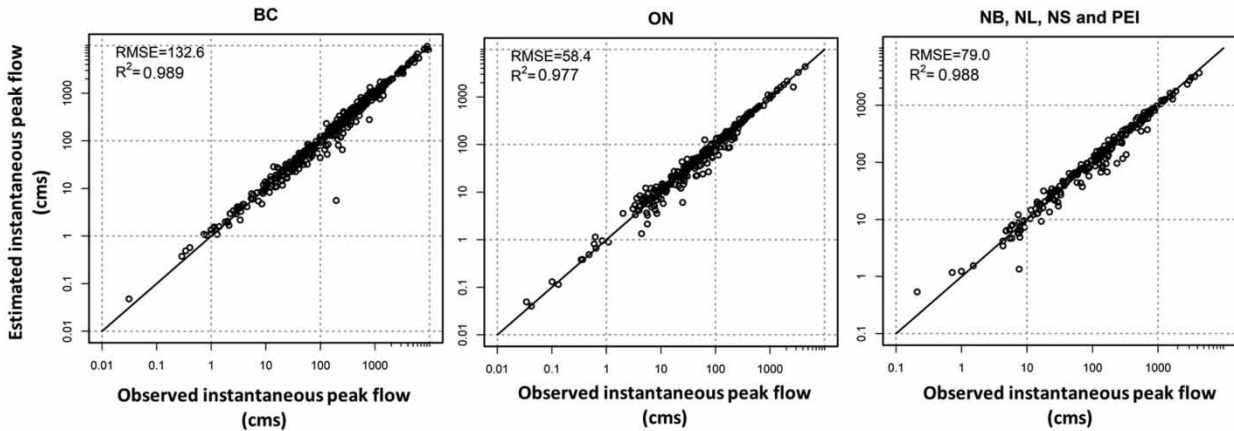
#### 4.1. Machine learning approach

As mentioned before, an ANN-based method (M5) was also considered apart from the conventional and newly introduced nonlinear regression type methods. It is well known that machine learning methods require larger data sets for proper training purposes. Therefore, these methods were tested considering three data sets, i.e., IPFs and MDFs from all natural stations from BC, ON, and four Atlantic provinces (NB, NL, NS, and PEI) together in order to carry out relatively an unbiased assessment.

For developing ANN-based models, various network structures were evaluated for the three selected data sets considering 70% of data for training and 30% for validation in order to optimize the network structure and controlling parameters. It is generally believed that there is a tradeoff between accuracy and complexity of the network for a given problem (Maier *et al.* 2010). For the analyzed data sets, it was found that ANN structures with several hidden layers and neurons did not result in better performing models compared to simpler options. For ON, a structure with two hidden layers respectively containing nine and three neurons led to a better performing model. For BC and Atlantic provinces together, a structure consisting of two hidden layers respectively containing seven and five neurons resulted in a better model. After conducting training and validation, the optimal model was applied to the entire data set and the comparison of observed and estimated IPFs is provided in Figure 8. For BC and ON, the performance metrics are quite comparable to those provided in Figure 4 for methods M1–M4.



**Figure 7** | AICc-based rankings of methods M1–M4 for all P&Ts and watersheds, with natural flow regimes. Proportions of watersheds/stations with the associated best method at the first, second, and third place are shown.



**Figure 8** | Comparison of the POR highest observed and estimated (using method M5) IPFs for BC, ON, and together for NB, NL, NS, and PEI. Diagonal line represents the line of perfect match and the values of two performance measures (RMSE and  $R^2$ ) are indicated inside each panel.

For Atlantic provinces, the simpler structure performs reasonably well, with RMSE ( $R^2$ ) of 79 (0.988), compared to 73.9 (0.984), 62.3 (0.988), 89.2 (0.975), and 69.0 (0.985), respectively, for models M1–M4.

#### 4.2. Machine learning-based fusion modeling

To synthesize outputs from multiple methods, one can exploit an ensemble modeling framework and can combine outputs using several different approaches, e.g., simple averaging, weighted averaging, optimal combination, or Bayesian averaging.

In this study, machine learning-based fusion modeling was used to synthesize outputs of multiple methods for the same three cases as considered in the above section, i.e., POR highest IPFs for BC, ON, and Atlantic provinces. In the machine learning area, this is also known as super learning. An ANN-based fusion model was developed using M1–M5 estimates as input and observed IPFs as target. After experimenting with several shallow and deep ANN structures and multiple training strategies, a network with a single hidden layer and 10 hidden neurons was adopted for Atlantic provinces. This led to  $RMSE = 54.8$ ,  $R^2 = 0.991$ , and  $MAPE = 5.33$ , which are slightly better than those of individual methods discussed in the above section. In the case of ON, a network with two hidden layers with respectively nine and three hidden neurons was adopted. This resulted in  $RMSE = 58.3$ ,  $R^2 = 0.979$ , and  $MAPE = 3.69$ . These measures are similar to those presented in Figure 4 for methods M1–M4 and in the above section for method M5. In the case of BC, the fusion modeling approach did not lead to any better results than those of methods M2 and M5.

## 5. CONCLUSIONS

Information on IPFs is often required to derive design values for sizing various hydraulic structures, such as culverts, bridges, small dams, and levees, in addition to informing hydraulic procedures of floodplain delineations, erosion risk management strategies, and water resources management-related activities. In this paper, most of the previous work on the estimation of IPFs from MDFs is reviewed in the historical context. In particular, Sangal's method (M1), which was proposed for Ontario watersheds, is given special attention. Three new methods (M2–M4) are proposed by generalizing Sangal's method in terms of statistical functional relationships. These methods and a machine learning-based method are evaluated using IPFs and MDFs from Canadian watersheds, with both natural and regulated flow regimes, and multiple performance measures. To select a best performing method from the group of four competing methods, a corrected form of the AIC-based model selection criterion, which is well-established in the literature, is used. The following main conclusions can be drawn from the results presented and discussed in this paper.

Through a detailed evaluation of Sangal's original method (M1), it was found that the parameter  $\alpha$ , which varies between  $\pm 0.5$ , tends to become negative with increasing drainage area for most watersheds in most of the P&Ts, with the majority of the values found below zero. This suggests that it is not reasonable to assume  $\alpha = 0$  and therefore full form of the method should be used in future studies. The simplified method that results by assuming  $\alpha = 0$  is applicable only for a small number of watersheds. Through statistical simulation based inferential statistics, it is possible to define a range where this parameter can be assumed equal to zero. This is a reasonable research topic for future studies.

By modeling POR highest IPF and associated MDFs, it was found that the method M2 performs relatively better than the other three methods for most of the P&Ts. For example, M2 was selected as the most suitable method based on AICc for AB, ON, NL, MB, NB, NS, NT, and YT, while M1 was selected for BC and SK for the case of natural stations. For regulated stations, the same conclusion holds. In contrast, when a similar evaluation was performed for individual watersheds within each P&T, other methods also emerged as the best performing ones based on the AICc. Overall, the performance of various methods appears to be linked with the size of the watershed, i.e., the spread of performance measures  $R^2$  and MAPE was found to be relatively large for smaller watersheds than for larger ones. In any given situation, it is also important to consider additional constraints. For example, the estimated IPF must be greater than or equal to the maximum value of the sequence of MDFs used in its estimation.

In addition to conventional methods and regression-based relationships, machine learning approaches were also evaluated for estimating IPFs from MDFs. For this evaluation, POR highest IPFs for BC, ON, and Atlantic provinces together were considered. These provinces have the greatest number of flood events, which was necessary for an unbiased evaluation. It was found that a carefully trained and validated machine learning model can also be used to estimate IPFs from associated MDFs. By investigating several ANN structures, it was also found that simpler structures can perform much better than their complex counterparts for the data sets considered in this study.

Since this study represents a multiple modeling situation, the fusion modeling technique from the machine learning area was invoked in order to synthesize a single output by combining outputs of multiple methods. This was explored by considering POR highest IPF for BC, ON, and Atlantic provinces together and training ANN-based fusion models. These models led to slightly improved estimates of IPFs for Atlantic provinces and comparable estimates for ON. In the case of BC, fusion modeling results were inconclusive. Conceptually, fusion modeling is similar to ensemble modeling, which is commonly used in hydrology and meteorology to combine and synthesize multiple outputs from an array of different candidate models or

simulations of the same model. Given the importance of this technique, additional investigations are required, especially to combine outputs of multiple machine learning models, statistical functional relationships, and other modeling approaches.

## ACKNOWLEDGEMENTS

This work was completed within the framework of National Research Council Canada's Climate Resilient Built Environment (CRBE) initiative, which is funded through Infrastructure Canada (INFC). The financial support of INFC and the leadership of CRBE are gratefully acknowledged. All data analyses were performed on the R computing platform. The helpful comments of three anonymous referees and editors are very much appreciated.

## DATA AVAILABILITY STATEMENT

All relevant data are available from an online repository or repositories. [https://wateroffice.ec.gc.ca/mainmenu/historical\\_data\\_index\\_e.html](https://wateroffice.ec.gc.ca/mainmenu/historical_data_index_e.html)

## CONFLICT OF INTEREST

The authors declare there is no conflict.

## REFERENCES

- Akaike, H. 1974 A new look at the statistical model identification. *IEEE Trans. Automat. Contr.* **19** (6), 716–723. doi:10.1109/TAC.1974.1100705002E.
- Anastasiadis, A. D., Magoulas, G. D. & Vrahatis, M. N. 2005 New globally convergent training scheme based on the resilient propagation algorithm. *Neurocomputing* **64**, 253–270.
- Bonakdari, H. & Zeynoddin, M. 2022 Chapter 5 – Goodness-of-Fit & Precision Criteria. In: *Stochastic Modeling: A Thorough Guide to Evaluate, Pre-Process, Model and Compare Time Series with MATLAB Software*. Elsevier, Amsterdam, Netherlands, pp. 187–264.
- Brewer, M. J., Butler, A. & Cooksley, S. L. 2016 The relative performance of AIC, AICc and BIC in the presence of unobserved heterogeneity. *Meth. Ecol. Evol.* **7**, 679–692.
- Burnham, K. P. & Anderson, D. R. 2002 *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd edn. Springer, New York, NY, USA.
- Burnham, K. P. & Anderson, D. R. 2004 Multimodel inference understanding AIC and BIC in model selection. *Sociol. Methods Res.* **33** (2), 261–304. doi:10.1177/0049124104268644.
- Chattopadhyaya, G., Midyaa, S. K. & Chattopadhyay, S. 2019 MLP based predictive model for surface ozone concentration over an urban area in the Gangetic West Bengal during pre-monsoon season. *J. Atmos. Sol.-Terr. Phys.* **184**, 57–62.
- Chen, B., Krajewski, W. F., Liu, F., Fang, W. & Xu, Z. 2017 Estimating instantaneous peak flow from mean daily flow. *Hydrol. Res.* **48** (6), 1474–1488. doi:10.2166/nh.2017.200.
- Creager, W. P., Justin, J. D. & Hinds, J. 1945 *Engineering for Dams*, Vol. 1. John Wiley and Sons, New York, NY, USA.
- Dastorani, M. T., Koochi, J. S., Darani, H. S., Talebi, A. & Rahimian, M. H. 2013 River instantaneous peak flow estimation using daily flow data and machine-learning-based models. *J. Hydroinform.* **15** (4), 1089–1098.
- ECCC 2023 *National Water Data Archive: HYDAT*. Environment and Climate Change Canada, Government of Canada. Available from: <https://www.canada.ca/en/environment-climate-change/services/water-overview/quantity/monitoring/survey/data-products-services/national-archive-hydat.html>.
- Ellis, W. & Gray, M. 1966 Interrelationships between the peak instantaneous and average daily discharges of small prairie streams. *Can. Agri. Eng.* February 1–25.
- Fill, H. D. & Steiner, A. A. 2003 Estimating instantaneous peak flow from mean daily flow data. *J. Hydrol. Eng.* **8**, 365–369.
- Helsel, D. R. & Hirsch, R. M. 2002 *Statistical Methods in Water Resources*. U. S. Geological Survey, VA, USA, p. 524.
- Hurvich, C. M. & Tsai, C. L. 1989 Regression and time series model selection in small samples. *Biometrika* **76** (2), 297–307.
- IPCC 2013 In: *Climate Change 2013. The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., Allen, S. K., Boschung, J., Nauels, A., Xia, Y., Bex, V. & Midgley, P. M. eds.). Cambridge University Press, Cambridge, UK and New York, NY, USA.
- Jimeno-Sáez, P., Senent-Aparicio, P., Pérez-Sánchez, J., Pulido-Velazquez, J., and Cecilia, D. & M, J. 2017 Estimation of instantaneous peak flow using machine-learning models and empirical formula in Peninsular Spain. *Water* **9**, 347. doi:10.3390/w9050347.
- Khaliq, M. N. 2023 Estimation of instantaneous peak flows in Atlantic Canada. In *Annual Conference of the Canadian Society for Civil Engineering*, Moncton, NB, Canada.
- Langbein, W. 1944 Peak discharge from daily records. *Water Resour. Bull.* August, 145.
- Lantz, B. 2015 *Machine Learning with R*, 2nd edn. Packt Publishing, Birmingham, UK.
- Maier, R. H., Jain, A., Graeme, C. D. & Sudheer, K. P. 2010 Methods used for the development of neural networks for the prediction of water resource variables in river systems: Current status and future directions. *Environ. Model. Softw.* **25**, 891–909.

- Neill, C. R. 1986 Unusual Canadian floods and the Creager diagram. *Can. J. Civ. Eng.* **13**, 255–257.
- R Core Team 2023 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.  
Available from: <https://www.R-project.org/>.
- Sangal, B. P. 1981 *A Practical Method of Estimating Peak From Mean Daily Flows with Application to Streams in Ontario*. National Hydrology Research Institute, Inland Waters Directorate, Ottawa, Canada, p. 247.
- Sangal, B. P. & Kallio, R. W. 1977 *Magnitude and Frequency of Floods in Southern Ontario*. Technical Bulletin Series No. 99, Inland Waters Directorate, Water Planning and Management Branch, Ottawa, Canada, p. 349.
- Singh, V. P. 1995 *Computer Models of Watershed Hydrology*. Water Resources Publications, Highlands Ranch, CO, USA.
- Singh, V. P. & Frevert, D. K. 2002 *Mathematical Models of Small Watershed Hydrology and Applications*. Water Resources Publications, Highlands Ranch, CO, USA.
- Teufel, B. & Sushama, L. 2021 2°C vs. high warming: Transitions to flood-generating mechanisms across Canada. *Water* **13** (11), 1494.
- Walpole, R. E., Myers, R. H., Myers, S. L. & Ye, K. 2012 *Probability and Statistics for Engineers and Scientists*. Prentice Hall, Boston, MA, USA.
- Watt, W. E. 1989 *Hydrology of Floods in Canada – A Guide to Planning and Design*. National Research Council Canada, Ottawa, Canada, p. 263.
- Wolfs, V. & Willems, P. 2014 Development of discharge-stage curves affected by hysteresis using time varying models, model trees and neural networks. *Environ. Model. Softw.* **55**, 107–119.

First received 15 July 2023; accepted in revised form 14 March 2024. Available online 17 April 2024