ARTIFICIAL NEURAL NETWORK-BASED FLOOD FORECASTING: INPUT VARIABLE SELECTION AND PEAK FLOW PREDICTION ACCURACY

EVERETT SNIEDER

A THESIS SUBMITTED TO THE FACULTY OF GRADUATE STUDIES IN PARTIAL FULFILLMENT OF THE REQUIREMENTS OF THE DEGREE OF MASTER OF APPLIED SCIENCE

GRADUATE PROGRAM IN CIVIL ENGINEERING

York University

Toronto, Ontario

August 2019

ABSTRACT

Floods are the most frequent and costly natural disaster in Canada. Flow forecasting models can be used to provide an advance warning of flood risk and mitigate flood damage. Data-driven models have proven to be suitable for flow forecasting applications, yet there are several outstanding challenges associated with model development. Firstly, this research compares four methods for input variable selection for data-driven models, which are used to minimize model complexity and improve performance. Next, methods for reducing the temporal error for data-driven flood forecasting models are investigated. Two procedures are proposed to minimize timing error: error weighting and least-squares boosting. A class of performance measures called visual measures is used to discriminate between timing and amplitude errors, and hence quantifying the impacts of each correction procedure. These studies showcase methods for improving the performance of flow forecasting models, more reliable flood risk predictions, and better preparedness for flood events.

DEDICATION

This thesis is dedicated in loving memory to Sarah Mason, comet chaser.

Illustration by Sarah

ACKNOWLEDGEMENTS

Foremost, I would like to thank my supervisor Dr. Usman T. Khan for his unceasing guidance and rich discussion, whether on the topic of cycling or fuzzy numbers. Also, for his patience and compassion during a time of tribulation.

Next, I would like to thank Dr. Kevin Gingerich for his valuable feedback and assistance, often on very short notice.

Thank you Rahma Shakir for providing preliminary work on input variable selection and useful feedback.

Thank you to Dr. Tom Gleeson, for granting me an office space among his research team at the University of Victoria.

Thank you to the Government of Canada and the Toronto and Region Conservation Authority for providing the data used in this research.

Lastly, thank you to York University and Enbridge Inc. for providing funding for this research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACF	Autocorrelation function
AIC	Akaike information criterion
ANN	Artificial neural network
BP	Backpropagation
CCF	Cross-correlation function
CNPS	Combined neural pathway strength
CVG	Coverage
DDM	Data-driven model
DFAA	Disaster financial assistance arrangements
EW	Error weighting
EWS	Early warning system
HM	Hydrograph matching
IO	Input omission
IVS	Input variable selection
KFCV	K-fold cross-validation
KS2	Two sample Kolmogorov-Smirnov test
LMBP	Levenberg-Marquardt backpropagation
LSB	Least-squares boosting
ML	Machine learning
MSE	Mean squared error
NSE	Nash-Sutcliffe efficiency
NSET	Nash-Sutcliffe efficiency timing
PC	Partial correlation
PD	Peak difference
PI	Persistence index
PMI	Partial mutual information
PRC	Precision
RMSE	Root mean squared error
SD	Series distance
SSE	Sum of squared error
VM	Visual measure

CHAPTER 1. INTRODUCTION

Floods are events that occur when water inundates regions that are typically dry. While floods are a natural process, they are characterized as a natural disaster when their impacts are detrimental to public health. This chapter provides a synopsis of flood disasters in Canada, overviews flood types and terminology, and provides background on the various types of models used for forecasting flood events. Following this, the chapter outlines the specific research objectives of the thesis as well as an outline of the thesis.

1.1 Flood disasters in Canada

Foods are the most frequent and costly natural disaster in Canada (Public Safety Canada n.d.). According to a report published by the Office of the Parliamentary Budget Officer, flood events cause an estimated $2.43 billion in damage nationwide, which is approximately half of the estimated cost for all types of extreme weather events (Office of the Parlimentary Budget Officer 2016). The Disaster Financial Assistance Arrangements (DFAA), a federal program that provides financial assistance following disasters, contributes $0.673 billion annually for flooding, roughly three quarters of their annual contributions for extreme weather events (Office of the Parlimentary Budget Officer 2016).

There are 6517 recorded flood events between 1970-2014 that received a DFAA payment; however, the actual number of floods is larger as it includes events that do not receive federal assistance or go unrecorded altogether. While flood events are frequent, much of the total nationwide flood damage is typically attributable to a few large events each year. Two such events occurred in 2013, when the Bow River and Don River experienced severe flooding. These two floods are summarized in Table 1-1 below. Images of the Bow River and Don River flooding below show the extent of the area affected (Andy Clark 2013, Liem Vu 2013). Both photos are of areas in the downtown core of each respective city.

Table 1-1: Comparison of Bow River and Don River 2013 major flood events.

	Bow River	Don River
City affected	Calgary	Toronto
Event date	June 19 – 28, 2013	July 8, 2013
Precipitation	126 mm	75 – 150 mm
Fatalities	4	0
Estimated total cost	$2,715,742,000	$940,000,000

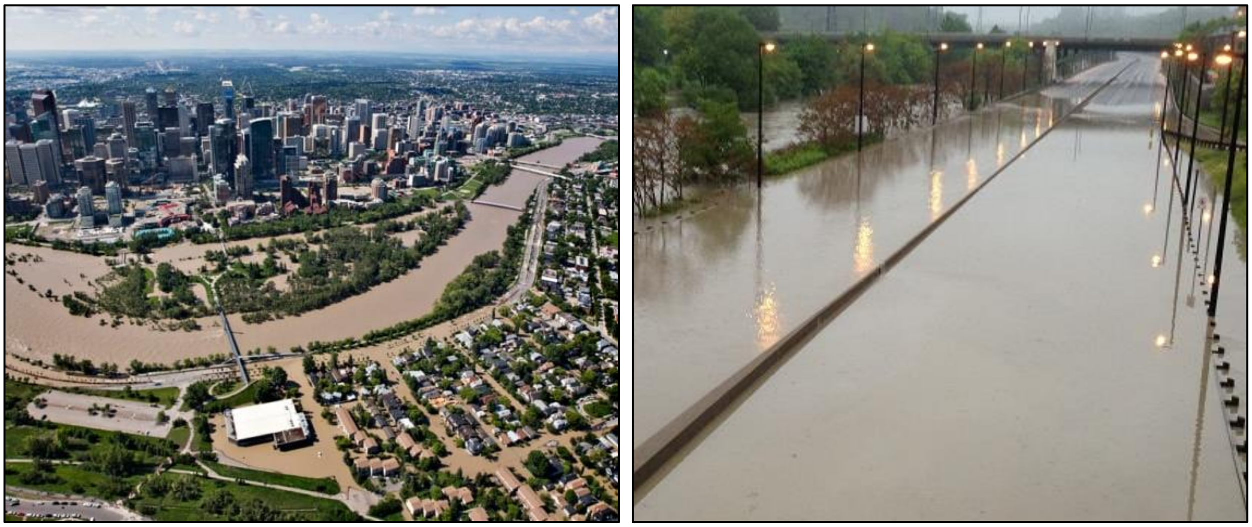

Figure 1-1: Left: major 2013 flood event in the Bow River; right: major 2013 flood event in the Don River.

In order to minimize the damage causes by flooding, it is important to first understand the underlying hydrological processes of these disasters. The following section provides a brief flood taxonomy and outlines some of the hydrological processes and terminology related to riverine flooding.

1.2 Flood drivers and hydrological processes

There are a variety of different flood types (i.e., riverine, pluvial, coastal, and groundwater flooding), which are often driven by one or several hydrological processes (i.e., rainfall, storm surges, and ice blockages in rivers). The two major flood events featured in the previous section, and all the flooding included in this research, is classified as riverine flooding. Riverine flooding occurs when the water level in a river rises and inundates the surrounding area, which is typically caused by high intensity and high volumes of precipitation.

Hydrographs depict changes in water flow or elevation as a function of time. They typically begin when flow levels increase above baseflow and end when the flow level returns to baseflow; baseflow is the regular streamflow, without contributions from rainfall events. Rainfall events are illustrated using hyetographs, which show precipitation depth as a function of time. Figure 1-2 contains a generic hyetograph (plotted along a reversed y-axis) and hydrograph, broadly illustrating the relationship between precipitation and streamflow. Precipitation that falls within a watershed becomes infiltration and runoff, runoff travels downstream and eventually enters a stream, causing water levels to increase. The leftmost portion of the hydrograph is referred to as the rising limb, as indicated in Figure 1-2. As precipitation slows or ceases, so does the runoff being generated. Water levels in the stream reach a peak value before decreasing; this portion of the hydrograph is called the receding (or falling) limb. The time interval between the start of the rising limb and the peak level is called the time to peak (National Research Council 2008). Similarly, time interval between the center of mass of precipitation and water level is called the lag time (National Research Council 2008). There is also a lag time between upstream hydrographs and downstream hydrographs, as the upstream flow contributions typically cause the downstream level to peak at a later point in time.

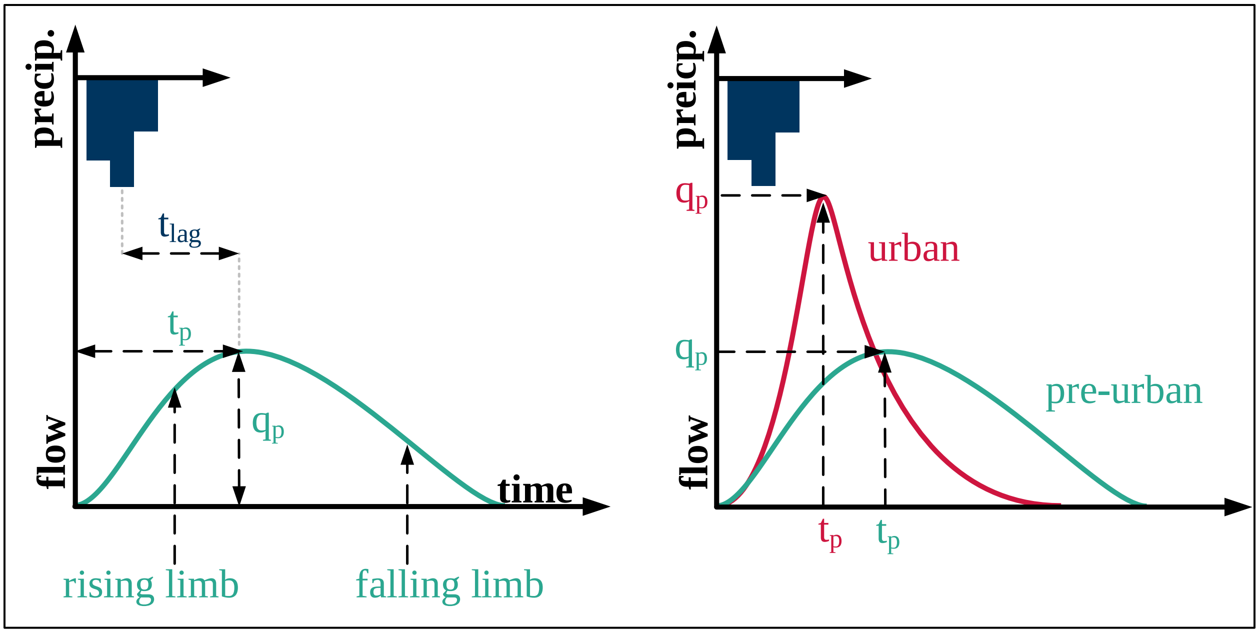

Figure 1-2: Left: example of time lag (tlag), time to peak (tp), rising and falling limbs; right: comparison of hydrographs for urban and pre-urban watersheds.

As mentioned above, riverine flooding is generally caused by large volumes of precipitation. Precipitation generates runoff, which in turn causes the water elevation in the receiving stream to rise. Flooding occurs when the water level exceeds a specific threshold value; high water levels are associated with a specific flood frequency (i.e., the 100-year flood has a corresponding flood plain). Since precipitation and upstream flow stations peak earlier than downstream stations, they provide advance notice of high water levels downstream, thus making suitable predictors for flood forecasting models.

1.3 Trends in flood severity

Roughly half of the world's population live in urban areas, a figure that is expected to continue to increase (Wilby 2007). The severity of flooding is exasperated by urbanization (Jongman 2018). One effect of urbanization is an increase in impermeable surfaces, which reduces the capacity for water to infiltrate, and instead, increases runoff generated (National Research Council 2008). The effects of urbanization on a downstream hydrograph are illustrated in Figure 1-2, in which the urbanized hydrograph has a shorter time to peak and higher peak flow compared to the pre-urban conditions. This translates to floods that are more severe and occur more rapidly. Urbanization is also associated with increased population density, meaning that flooding impacts a greater number of people (Jongman 2018). Climate-change projections anticipate more frequent extreme rainfall events, which will produce more severe and frequent flood events (Wilby 2006, 2007).

As conditions are trending to produce more severe flooding, methods for mitigating flood damage are advancing. Some emerging approaches for reducing flood damage include leveraging social media, improved urban planning practices, and improved flood forecasting models; the following research is focused on the final point (Jongman 2018).

1.4 Flow forecasting models

Hydrological models can be used to forecast water levels, providing an advanced warning of flood risk. The following section provides an outline of the conventional, physically-based approach to hydrological modelling, as well as the data-driven modelling (DDM) approach, which has been gaining popularity throughout the last four decades. The purpose of the following research is not to compare the performance of physically-based and data-driven models; although, the research makes comparisons between characteristics of physically-based and data-driven models (i.e., data requirements, interpretability, etc.). Therefore, it is important to provide a brief background on the two distinct modelling approaches.

1.4.1 Physically-based flow forecasting models

Physically-based (sometimes called knowledge-based) flow forecasting models attempt to simulate hydrological processes in order to predict downstream flow conditions. Due to the complexity nature and incomplete understanding of stormwater science, these models grossly oversimplify spatial and temporal variables (National Research Council 2008). Nevertheless, by calibrating uncertain model parameters, these models are capable of reasonably producing accurate downstream predictions.

The high data requirements, time consuming model development, assumptions, and oversimplifications are all deterrents for using such models. For these reasons, data-driven approaches are gaining popularity. Unlike physically-based approaches, DDMs can be calibrated to predict downstream conditions without preliminary assumptions or physical description of the hydrological system.

1.4.2 Data-driven flow forecasting models

While physically-based models rely on expert knowledge to understand the relationship between hydrological variables, DDMs make connections based on a mathematical assessment of the data. While an understanding of hydrology is helpful for developing such models, DDMs do not attempt to model hydrological processes. As a result, they have relatively low data requirements, low cost, and are simple to develop; yet, are capable of outperforming physically-base models, which has contributed to their increasing popularity among hydrologists (ASCE 2000a, Maier et al. 2010, Abrahart et al. 2012, Mosavi et al. 2018).

Despite their widespread use, there are several outstanding challenges associated with developing and interpreting flow forecasting DDMs (Abrahart et al. 2012). The following section outlines some of these challenges, followed by the objective of this research.

1.4.3 Limitation of current data-driven modelling approaches

One important stage of model development for ANNs is input variable selection (IVS) which is the process of selecting the best inputs for predicting output and achieving the highest overall predictive performance. While an understanding of hydrology can provide a modeller with a general sense for which inputs make good predictors, including inputs that are not relevant or have too much interdependency can hinder model performance (May et al. 2011). There are a wide variety of studies that develop or propose different IVS methods, however in cases where IVS is not a primary focus of the study, it is largely overlooked and researchers rely on expert knowledge or simple linear methods to infer input usefulness; input usefulness being a combination of maximizing similarity to the output while minimizing redundancy with the other inputs (May et al. 2011, Abrahart et al. 2012). Studies that develop IVS methods often rely on synthetic datasets in which candidate inputs are either useful or non-useful (Sharma 2000, May et al. 2008b). Hence there is a clear need to refine and compare existing IVS methods on real data, for their ability to identify useful inputs and determine the optimum number of inputs (Abrahart et al. 2012). Doing so will help create more accurate flow forecasting models to help with flood management and preparedness. This topic is explored in Chapter 2, which compares four distinct IVS methods.

Since real data is being used, the collective usefulness of a selection of inputs is determined by measuring the performance of the ANN they inform. It is considered good practice to use several difference performance measures for assessing ANNs (Maier et al. 2010, Bennett et al. 2013). There are a wide variety of different performance measures for flow forecasting (Bennett et al. 2013). The most popular performance measures are based on the squared residuals of the model, such as Nash-Sutcliffe efficiency (NSE) and root mean squared error (RMSE). Despite their popularity, such measures are widely criticized, partly due to their insensitivity to characteristics such as seasonality in the observed data or timing errors between the observed and predicted timeseries (Gupta et al. 2009, Ehret and Zehe 2011). For instance, in a seasonal watershed, predictions with a timing error equal to the forecast lead time may exhibit an exceptionally high NSE value; however, this model has no predictive value.

This becomes especially important given that timing errors are frequently observed in ANN predictions (Conway et al. 1998, Abrahart et al. 2010). This is partly because ANNs typically use mean squared error (MSE) or similar as a cost function for calibration. This issue is especially important for models that inform early warning systems (EWS), where forecast timing error translates to a reduced amount of time to implement flood management measures. Investigating the timing error challenge is the central focus of Chapter 3, which evaluates two correction procedures for modifying the ANN calibration procedure to reduce timing error. A special class of performance measures, called visual measures (VMs), are used to assess the impact of each correction procedure. VMs are intended to mimic a hydrologist's comparison of two hydrographs by disentangling error into amplitude and timing components. Three different VMs are used and compared with each other and how well they agree with a visual assessment of two hydrographs.

1.5 Research objectives

The following sections list the specific research objectives categorized by topic: 'IVS research' (Chapter 3) and 'Peak flow research' (Chapter 4).

1.5.1 IVS research objectives

i. Refine and develop novel improvements to existing IVS methods

ii. Conduct a comprehensive comparison of model-based and model-free IVS methods

iii. Evaluate the efficacy of a termination criteria for establishing the number of model inputs

Meeting the objectives outlined above will equip model developers with more tools and knowledge for choosing model inputs. In model applications, choosing optimum model inputs improves model performance and reduces data requirements.

1.5.2 Peak flow performance research objectives

i. Evaluate and apply methods for improving peak flow accuracy

ii. Quantify improvement using specialized performance measures

iii. Compare continuous and event-based performance assessment

Meeting the objectives outlined above will provide modellers with additional tools improving the peak flow performance of flow forecasting models. The benefits of improving model performance assessment applies to all types of flow forecasting models, whether data-driven or physically-based.

1.5.3 Thesis outline

The study regions (the Bow and Don Rivers) are summarized in Chapter 2 below, which contains a summary of the watersheds, hydrometeorological data, data preprocessing, and a description of model inputs and outputs. Next, Chapter 3 presents the methods, results, and discussion for the IVS research. Chapter 4 contains the methods, results, and discussion for the peak flow correction procedures and model performance assessment. The research findings are summarized in the conclusion and future research opportunities are outlines. The appendices contain additional results for each of the studies and pseudocode for IVS methods.

CHAPTER 2. STUDY AREA

This chapter provides descriptions of the two regions studied in this thesis, the Bow (Upper and Central) and the Don River watersheds. Each watershed is described firstly on their physical characteristics, followed by statistical summary of the available hydrometeorological variables, and finally a description of the hydrometeorological variables used as inputs and outputs for the models.

2.1 Description of watersheds

This research features two distinct watersheds, the Bow River (Upper and Central) and the Don River, which are located in western and central Canada, respectively, as indicated in Figure 2-1. The Bow River watersheds have headwaters fed by the Rocky Mountains and contain predominantly natural and agricultural land uses. In comparison, the Don River watershed is heavily urbanized; large regions of impermeable surfaces and stormwater management infrastructure is present throughout the basin. Hydraulic infrastructure is present along both rivers, which is evidenced in the downstream hydrographs that have characteristic and unnatural receding limbs, which are attributable to the operation of a dam. The presence of infrastructure is also confirmed by available documentation for each river. A comparative summary of the watersheds' attributes is provided in Table 2-1 directly below.

Table 2-1: Bow and Don River watershed characteristics.

	Upper and Central Bow	**Don**
Governing agency	Bow River Basin Council	Toronto and Region Conservation Authority
Size (km^2)	16 000	360
Major land uses	Agricultural, natural, urban	Urban
Timeseries length at target station	10 years	10 months

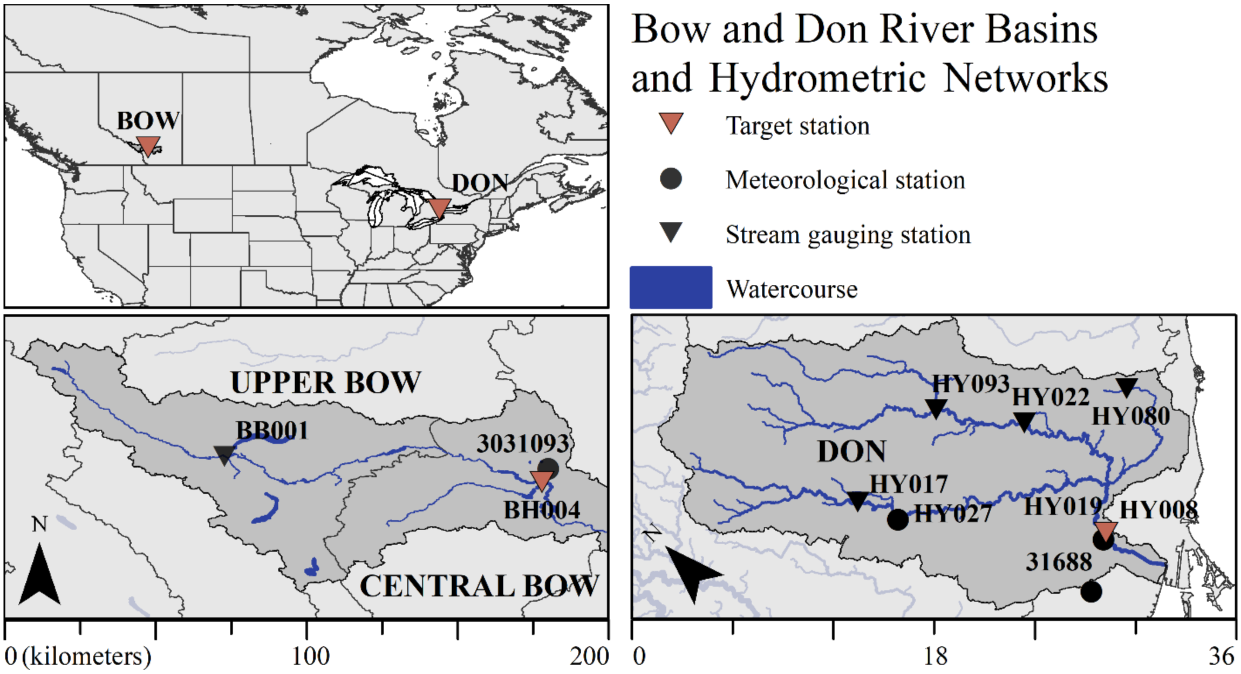

Figure 2-1: The location of the Bow River (bottom-left) and Don River (bottom-right) watershed systems in Canada; the target stations, and upstream meteorological and stream gauging stations are identified within each watershed.

The target stations were identified as suitable locations for forecasting since both have a history of flooding in recent years (namely in 2013).

2.2 Data preprocessing

Both the Bow and Don River datasets were relatively complete, containing few missing values. Timesteps at which any input variable contains a missing value are removed from the dataset; in other words, no missing data is interpolated. Data from November to April and November to December were removed from the Bow and Don rivers, respectively, due to snow or ice conditions.

Data is also automatically normalized as it is fed into the ANN in MATLAB, which is typical for such models. As such, differences in magnitudes between raw variable values are not a concern.

2.3 Statistical summary of variables

This section provides a brief statistical analysis of the meteorological timeseries used as input and output variables for the ANNs studied in this thesis research. Basic statistics for each of the hydrometeorological variables are listen in Table 2-2 to Table 2-4 below.

Table 2-2: Statistical summary for hydrometeorological timeseries (1-hour timestep) for the Don River watershed.

Stations	Units	Mean	Max.	Min.	Med.	90th P.	99th P.	Max. CCF	Max. CCF lag
HY019_WL_Mean	[m]	77.62	79.21	77.51	77.58	77.76	78.15	1.00	0
HY017_WL_Mean	[m]	180.79	182.30	180.67	180.75	180.93	181.40	0.76	1
HY093_WL_Mean	[m]	7.45	8.41	7.13	7.43	7.56	7.74	0.28	1
HY080_WL_Mean	[m]	150.67	152.82	150.59	150.64	150.73	151.14	0.72	2
HY022_WL_Mean	[m]	118.05	119.02	117.97	118.02	118.15	118.50	0.85	1
HY008_Precip_Sum	[mm]	0.08	18.30	0.00	0.00	0.00	1.83	0.22	4
HY027_Precip_Sum	[mm]	0.05	27.60	0.00	0.00	0.00	1.00	0.08	4
GC31688_Temp_Mean	[°C]	12.24	34.90	-24.40	13.90	25.50	30.60	0.07	-36

Table 2-3: Statistical summary for hydrometeorological timeseries (1-day timestep) for the Bow River watershed.

Stations	Units	Mean	Max.	Min.	Med.	90th P.	99th P.	Max. CCF	Max. CCF lag
BH004_WL_Mean	[m]	1.28	2.77	0.92	1.23	1.64	2.08	1.00	0
BH004_WL_Min	[m]	1.26	2.47	0.75	1.21	1.61	2.05	0.99	0
BH004_WL_Max	[m]	1.31	3.13	0.93	1.26	1.68	2.12	0.99	0
BB001_WL_Mean	[m]	2.32	3.48	1.71	2.33	2.75	3.03	0.74	2
BB001_WL_Min	[m]	2.30	3.43	1.69	2.31	2.72	2.99	0.73	1
BB001_WL_Max	[m]	2.33	3.51	1.72	2.35	2.77	3.07	0.74	2
Calgary_Precip_Sum	[mm]	1.15	72.40	0.00	0.00	3.10	18.98	0.05	1
Calgary_Temp_Mean	[°C]	4.60	26.09	-31.42	5.56	17.15	22.19	0.24	-24
Calgary_Temp_Min	[°C]	-1.13	18.40	-33.90	-0.10	10.40	14.70	0.29	-18
Calgary_Temp_Max	[°C]	10.36	34.00	-30.10	11.30	24.22	30.70	0.17	-26

These values provide a sense for typical water levels the difference in elevation between typical flows and high flows (i.e., the 99th percentile flows). The maximum cross-correlation function (CCF) value and corresponding lag index are included, providing an approximate sense for the linear correlation between each variable and the output. The first station in each table, indicated by the italicized station name, is the target variable, hence the maximum cross-correlation (technically, the autocorrelation function (ACF), in this case) value of 1.00 at a lag of 0.

Table 2-4: Statistical summary for hydrometeorological timeseries (6-hour timestep) for the Bow River.

Stations	**Units**	**Mean**	**Max.**	**Min.**	**Med.**	**90th P.**	**99th P.**	**Max. CCF**	**Max. CCF lag**
BH004_WL_Mean	[m]	1.28	3.07	0.92	1.24	1.63	2.10	1.00	0
BH004_WL_Min	[m]	1.27	3.01	0.75	1.23	1.63	2.06	1.00	0
BH004_WL_Max	[m]	1.29	3.13	0.92	1.25	1.64	2.11	1.00	0
BB001_WL_Mean	[m]	2.32	3.51	1.69	2.33	2.75	3.03	0.73	7
BB001_WL_Min	[m]	2.31	3.50	1.69	2.33	2.74	3.03	0.73	7
BB001_WL_Max	[m]	2.33	3.51	1.70	2.34	2.76	3.04	0.73	7
Calgary_Precip_Sum	[mm]	0.29	72.40	0.00	0.00	0.40	6.34	0.02	7
Calgary_Temp_Mean	[°C]	4.60	33.15	-32.82	5.44	18.08	26.24	0.16	-96
Calgary_Temp_Min	[°C]	2.05	31.90	-33.90	2.70	14.60	24.60	0.17	-68
Calgary_Temp_Max	[°C]	7.03	34.00	-31.80	7.70	21.30	28.90	0.15	-96

The ACF and CCF are calculated for the target variable and between the target and upstream (non-autoregressive variables are also referred to as exogenous) variables, respectively. The ACF and CCF are useful for characterizing watersheds; they can be used to indicate properties such as seasonality (e.g., annual high and low flow periods) or approximate the lag time between variables (e.g., the delay between peak precipitation and water level). The ACF is the correlation (dependant variable) between a variable and itself at different lag times (independent variable). Similarly, the CCF is the correlation between a variable with a different variable, as a function of different lag times. These functions are often visualized using correlograms, that show lag times as the x-axis and the correlation as the y-axis.

In this study, calculating the ACF for flow in the Bow River watershed reveals annual seasonality (also called periodicity), whereas no seasonality is observed for the Don River. Seasonality was determined using 10 years worth of data for each watershed. However, for training the ANN, the timeseries length of the Don River dataset is limited for the target station: only 10 months of common data were available between the candidate inputs and output data. On the other hand, 10 years' worth of data were available for the Bow River target station.

The CCFs between the target station and other stations are used to approximate the lag times between the stations (Talei and Chua 2012). Ideally, upstream variables used as inputs should be lagged to reflect the actual travel time in the watershed. However, the lag time between stations is neither precise nor stationary. The lag time corresponding to the largest correlation provides a reasonable estimation for the lag time between stations. However, while developing the candidate input set, many lag times are used, which ensures that all potentially useful lag times are not omitted from the candidate set and because the cross-correlation only provides an estimate of the linear similarity between two stations and there many be non-linear similarities that are not captured. The approach described above is commonly used for identifying potentially useful inputs (cf. (Nanda et al. 2016, Tongal and Booij 2018)); however, the CCF is not recommended as a standalone IVS method because it does not consider interdependencies and non-linear relationships between variables (Bowden et al. 2005a).

The correlogram in shows how the hydrological system in the Bow River watershed is highly autocorrelated. Consequently, the model is prone to favouring autoregressive inputs, which contribute to timing error.

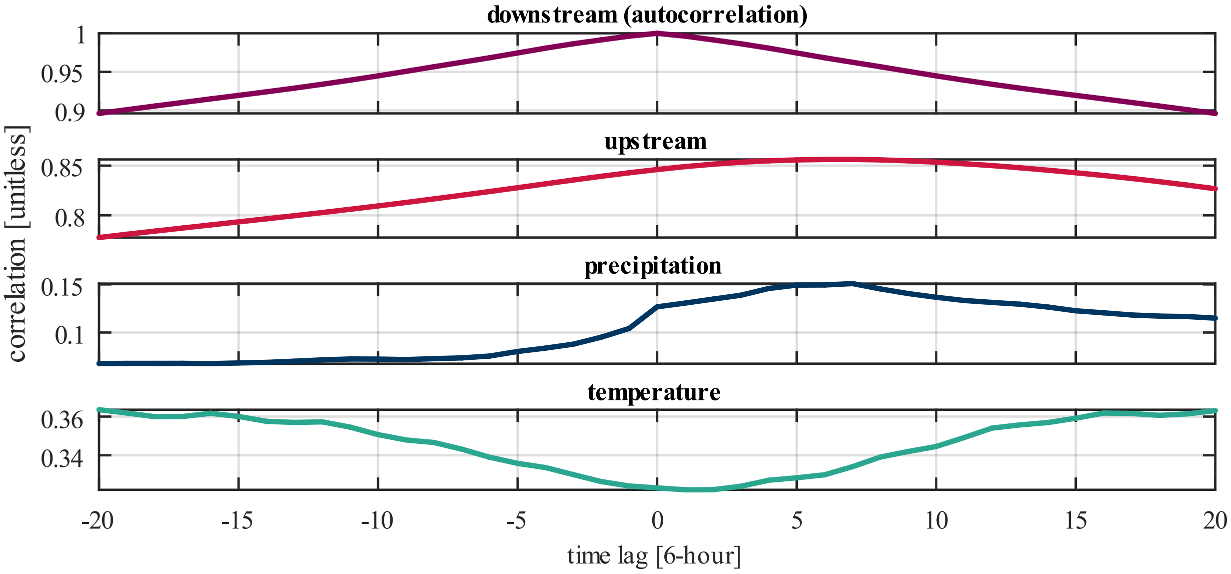

Figure 2-2: Autocorrelation (downstream) and cross-correlation (downstream and upstream) for ± 20 timesteps.

2.4 Description of input variables

The input sets for both studies are summarized in Table 2-5 and Table 2-6 below. The target stations for Bow and Don watersheds are BH004 and HY019, at various lead (forecast) times, as specified in the rightmost column in both tables. The number of times each input is lagged is specified in the 'Lag time' column. For example, each of the water level gauges for the Don River are lagged by 0, 1, 2, 3, 4, 5 hours in time, to inform distinct models with lead times of 1, 3, and 6 hours. The total number of inputs is the product of the number of distinct stations in each row and the number of time lags (e.g., for the Don there are 5 water level gages that each have 6 different lag times, for a total of 30 water level input variables).

All the data used in this research were collected and distributed by Environment Canada, the Water Survey of Canada, or the City of Calgary.

Table 2-5: Summary of all candidate input variables for the Don (hourly resolution) and Bow (daily resolution) River systems used in Chapter 3.

	Station ID	**Data type**	**Units**	**Data source**	**Lag times**	**Total inputs**	**Lead times**
Don River	HY017, HY019, HY022, HY080, HY093	Hourly water elevation	[m]	WSC[1]	0:1:5h	30	1, 3 & 6h
	HY008, HY027	Hourly precipitation	[mm]	WSC[1]	0:1:11h	24	
	31688	Hourly temperature	[°C]	Environment Canada	0:1:5h	6	
Bow River	BB001, BH004	Max, min, mean daily water level	[m]	WSC[1]	0:1:2d	18	1, 2 & 3d
	3031093	Cumulative daily precipitation	[mm]	City of Calgary	0:1:2d	3	
	3031093	Max, min, mean daily temperature	[°C]	City of Calgary	0:1:2d	9	

[1] Water Survey of Canada

The input sets in Table 2-5 comprise the candidate sets from which inputs are selected using IVS in Chapter 3. Removing non-useful inputs from the input set has numerous benefits such as improving performance, reducing computational demand, and reducing data costs; these benefits are discussed at length in Chapter 3.

Table 2-6: Summary of all the input variables for the Bow River system (6-hour resolution) used in Chapter 4.

	Station ID	**Data type**	**Units**	**Data source**	**Lag times** [6-hour]	**Total inputs**	**Lead times** [6-hour]
Bow River	BB001, BH004	Max, min, mean daily water level	[m]	WSC[1]	0:11	72	4
	3031093	Cumulative daily precipitation	[mm]	City of Calgary	0:11	12	
	3031093	Max, min, mean daily temperature	[°C]	City of Calgary	0:11	36	

[1] Water Survey of Canada

No input variable selection (IVS) is performed on the input set in Table 2-6; all 120 inputs are used in each model. While IVS could drastically reduce the computational effort and improve the precision of the calibration process, the intention is to not restrict the data available for calibration. For example, a neural network may favour autoregressive input variables by assigning high weights to their neural pathways; consequently, model-based IVS methods will identify these inputs as the most useful predictors. This is problematic, because while the model may produce a strong MSE value, models that underutilize upstream, exogenous data are more likely to exhibit a timing error. Therefore, not utilizing IVS ensures that all potential predictive inputs are available for model calibration.

CHAPTER 3. A COMPREHENSIVE COMPARISON OF FOUR INPUT VARIABLE SELECTION METHODS FOR ARTIFICIAL NEURAL NETWORK-BASED FLOW FORECASTING MODELS

3.1 Introduction

In coming decades, the risk of riverine flooding in urban areas is expected to increase, which is driven by factors such as climate change and rapid urbanization (Toronto and Region Conservation Authority 2018). One way that flood damage can be mitigated is with early flood warning systems which are used to provide advance warning of flood conditions to local authorities and floodplain occupants, reducing the risk of property damage and loss of life (Shrubsole et al. 1993, Yin et al. 2004). Flood warning systems typically utilize rainfall-runoff models to estimate future stage or discharge levels.

Historically, rainfall-runoff models have been developed based on physical processes. These models rely on the simplification of complex hydrological processes, which are highly nonlinear, and exhibit high spatial and temporal variability (ASCE 2000b, Wijesekara et al. 2012, Khan and Valeo 2016b). Throughout the past two decades, data-driven models (DDMs) have emerged as competitive alternatives to physically-based models for characterizing rainfall-runoff systems (ASCE 2000b, Dawson and Wilby 2001, Shrestha and Nestmann 2009). While physically-based models rely on principles of physics to describe a system, DDMs are based on mathematical relationships between data that characterize the system (Solomatine and Ostfeld 2008). Among data-driven methods, artificial neural networks (ANNs) are the most widely used for flow forecasting (Solomatine and Ostfeld 2008).

3.1.1 Artificial neural network-based flood forecasting

ANNs, which are inspired by biological neural networks, have a framework consisting of interconnected groups of input, hidden, and output nodes (see Figure 3-1). Each connection has a numeric weight, while hidden and output nodes both have a numeric bias and an activation function (tan-sigmoid and linear, respectively). Weights and biases are typically initialized to a random starting point, then adjusted based on an objective or cost function (such as the sum of squared error). Recent research in this field has largely focused on optimization algorithms that are used to train (or calibrate) the models, such as the use of swarm optimization and evolutionary algorithms to boost model performance to improve the accuracy of future predictions of stage or discharge in rivers (Meshram et al. 2018, Maier et al. 2018). However, there are fundamental aspects of ANN model development that are still widely overlooked, including: optimizing the architecture or the structure of the networks (i.e., the number of layers, and nodes in the hidden layers); the uncertainty associated with the amount of data used for training versus validation, or testing the models; and the selection of the best inputs needed for optimal model performance (May et al. 2011, Abrahart et al. 2012). This last component is the focus of the present research.

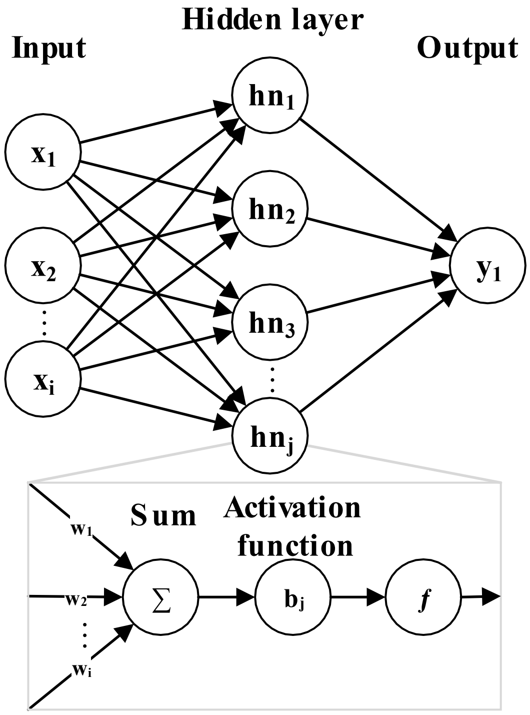

Figure 3-1: A schematic showing the components of a typical feed-forward multi-layer perceptron ANN model including the inputs (x_i), nodes in the hidden layer (hn_j), the output (y), the weights (w_i), the biases (b_j), and the activation function (f).

3.1.2 Overview of input variable selection

While ANN models have demonstrated their suitability for modelling rainfall-runoff systems, the selection of input variables (a process commonly referred to as Input Variable Selection, IVS) consistently receives little attention (Maier and Dandy 2000, Maier et al. 2010, May et al. 2011). This may be attributed to a reliance ANN training procedures for distinguishing useful from non-useful inputs, however, with no consideration is given to model complexity, learning difficulty, and performance (May et al. 2011). IVS methods are typically used to identify the most useful inputs from a larger set of candidate input variables; where usefulness is defined as having maximum relevancy to the output while minimizing redundancy with other candidate inputs (May et al. 2011). Reducing the number of model inputs is important for minimizing computational demand, reducing output variability caused by local minima on the error surface, and informing the behaviour of the physical system (Šindelář and Babuška 2004, Bowden et al. 2005a, May et al. 2008a).

Many applications of ANNs in hydrology do not describe a systematic IVS process, or rely on methods such as *a priori* knowledge of the system, trial-and-error, or linear cross-correlation (Maier and Dandy 2000, Bowden et al. 2005a, Abrahart et al. 2012). However, each of these approaches has limitations. While expert knowledge is convenient for model development and validating model behaviour, such knowledge is not dependably available and may not be suitable for identifying information such as interdependencies between hydrological variables, which is useful for IVS. Trial-and-error approaches, where a brute-force method is used to determine the best inputs, are computationally intensive, especially for ANN systems with a large number of candidate inputs (May et al. 2011). Lastly, linear cross-correlation, which is the most commonly used data-driven IVS approach, is limited to identifying discrete, linear relationships between the output and individual candidate inputs (ASCE 2000b, Abrahart et al. 2008, Nanda et al. 2016). While methods such as cross-correlation may be useful for reducing the number of inputs included in the set of candidates by providing modellers with the approximate lag times between monitoring points, it is unable to capture the interdependencies, redundancies, and nonlinearities typical of hydrological systems (Abrahart et al. 2012). For example, two rain gages situated close to each other may both have a strong correlation with a downstream flow gage, however, they be very similar and contain highly redundant information; if both are used as inputs, they may add unnecessary model complexity or increase learning difficulty. Therefore, there is a clear need for more robust IVS methods that do not rely on *a priori* knowledge or assumptions about the system, are computationally inexpensive, and can characterise nonlinear and interdependent relationships between candidate inputs. IVS has been the focus of multiple review papers that provide a comprehensive overview of its taxonomy and methods in water resources modelling (Bowden et al. 2005a, Maier et al. 2010, May et al. 2011). IVS methods can be grouped into two broad categories: model-free methods and model-based methods, both of which are described below, along with the four IVS techniques (Partial Correlation, Partial Mutual Information, Input Omission, and Combined Neural Pathway Strength) that are the focus of this research. Following this review, the overall specific objectives for this chapter are presented.

3.1.2.1 Model-free methods

Model-free IVS methods do not rely on a pre-existing model (Bowden et al. 2005a, May et al. 2011). Most IVS methods belong to this classification, and include common methods such as *a priori knowledge* and linear correlation. The first IVS method examined in this research is Partial Correlation (PC), which uses the partial correlation as a selection criterion; partial correlation is the linear correlation between two variables, controlling for the linear effects of one or several other variables. Unlike most correlation-based methods which do not consider linear redundancies between selected variables, the PC criterion reduces the likelihood of linear redundancy between input variables by iteratively selecting inputs from the candidate set, and controlling each new selection with the inputs that have already been selected (May et al. 2011). The PC method iteratively selects inputs using a forward-selection algorithm, in which the partial correlation between candidate inputs and the target variable is calculated at each step, where the input selected depends on the previous iteration (Sharma 2000, May et al. 2008b, He et al. 2011). This approach is criticized by May et al. (2008), who stated that linear, model-free methods are not suitable for non-linear models such as ANNs. He et al. (2011) demonstrated that, while inferior to its Partial

Mutual Information (PMI) based counterpart described below, PC is reasonably capable of identifying useful inputs.

The second model-free method examined in this research is PMI. The PMI criterion is calculated similarly to the PC criterion, by calculating the similarity between two variables, considering the effects of a control set. However, the PMI criterion is non-linear and calculated based on non-parametric kernel regression and kernel density estimates, which are discussed in greater detail in section 3.2. Like PC, PMI is used in a forward-selection algorithm, since each selection depends on the previous one. PMI-based IVS was first used by Sharma (2000), who used a bootstrapping algorithm to check for statistical significance of each PMI value, which is used as the termination criterion for the algorithm. Bowden et al. (2005b, 2005a) apply PMI to predict river salinity, using a two-step algorithm: where PMI is firstly used to select appropriate lagged inputs for each monitoring station, then selects from the reduced set of candidate inputs. This method is well suited for large candidate input sets where a large number of candidate inputs are available (Bowden et al. 2005b, 2005a). This application of PMI correctly identified all of the relevant input variables for a synthetic dataset (Bowden et al. 2005b, 2005a). May et al. (2008b, 2008a) evaluated different termination criteria for the PMI selection algorithm for synthetic and water quality datasets. In addition to the bootstrap-based approach, May et al. (2008b, 2008a) demonstrated the use of the Akaike Information Criterion (AIC) as an effective termination criterion for the PMI algorithm. Lastly, He et al. (2011) also demonstrated the capabilities of PMI using the AIC-based termination criterion, for selecting input variables for an ANN used to predict stormwater runoff quality parameters. Both PC and PMI are typically found to be capable of identifying useful inputs, producing benefits attributable to IVS such as improved performance and reduced complexity.

3.1.2.2 Model-based methods

Model-based IVS methods select inputs based on a calibrated model. After inputs are selected, models are recalibrated using a subset of the original inputs (May et al. 2011). The main limitations of this method are its high computational demand and its dependency on parameters that are decided prior to the model being trained (also called hyperparameters). For example, the inputs selected by a model-based method may depend on the number of hidden neurons chosen for the ANN. While hyperparameter optimization is not a focus of this research, ensemble modelling is used to capture some of the uncertainty associated with the selection of model parameters. Specifically, in this research, K-Fold Cross Validation (KFCV) and multi-start are used to generate ensembles, which is discussed in greater detail in section 3.2.1.

Input Omission (IO), a type of model-based IVS, estimates input usefulness by iteratively examining model behaviour following the omission of an input from the full set of inputs with which the ANN was trained. IO identifies useful and non-useful inputs-based on the significance (or insignificance) of the error caused by the omission of a certain input (Setiono and Liu 1997). Setiono and Liu (1997) use IO to select inputs for a classifier ANN; inputs are iteratively removed from the selected set-based on the input corresponding to the smallest decrease in accuracy when omitted from the model. The model is retrained at each iteration and the algorithm terminates once a maximum decrease in accuracy is reached. Next, Abrahart et al. (2001) use input omission (referred to as saliency analysis in their paper) to explain ANN behaviour and identify important inputs. Timeseries plots are generated from IO to infer the effect of each input on the model output (e.g., how omitting precipitation impacts the rising limb of the modelled hydrographs). This analysis provided defence against criticism of ANNs as being ‘black box models’, in cases where the data-driven input omission agrees with expert knowledge. For example, omitting an input that is known to be related directly to the output produces a decrease in model performance.

Lastly, Combined Neural Pathway Strength (CNPS) is another model-based method that uses a linear approximation of the total neural pathway strength of each input to estimate its usefulness. Generally speaking, the strength of a neural pathway is defined as the absolute magnitude of the weights associated with each input: the higher the magnitude of the weight, the stronger or more relevant a particular input is in predicting the desired output. Nash et al. (1997) demonstrate an early use of this idea, where the neural pathway strength is calculated as the relative, absolute matrix multiplication between the first and second sets of weights of a 3-layer ANN. The CNPS of each input is expressed as a percentage of the overall strength of all inputs. The approach is found to improve classification accuracy across several synthetic and real-world datasets. More recently, Duncan (2014) adopted a similar IVS approach, in which the CNPS is approximated as the simple matrix multiplication between the two weight matrices of a 3-layer ANN. The sign of the CNPS is used to indicate whether the input has an excitatory (positive correlation with output) or inhibitory (negative correlation with output) effect on the output. This method relies on a model ensemble framework, in which a distribution of CNPS values is calculated for each input: if the distribution is centred around CNPS values of 0, the input is deemed non-useful, whereas if the distribution is centred at higher magnitudes of CNPS values it is deemed to be useful (Duncan 2014). While this approach relies on simplifications and lumping ANN parameters (e.g., ignoring the effects of the activation functions and biases within the ANN), it has demonstrated capable performance for identifying useful inputs (Duncan 2014). This approach has been used in several recent studies where the use of Duncan's CNPS method resulted in unchanged or improved model performance for applications in water quality prediction, forest fire extent prediction, flood forecasting, and fever outbreak prediction (Duncan 2014, Khan et al. 2018, Laureano-Rosario et al. 2018).

The four IVS methods are summarized in Table 3-2 in the following Methods section. Each of the IVS methods described above, whether model-based or model-free, are conceptually distinct in their approach to selecting the most useful inputs from a set of candidate inputs. An important factor for each method is the termination criterion, which is used to determine whether or not a candidate input is significant (May et al. 2008a). For example, PC uses an increase in the complexity-based AIC criterion as the termination criterion for the selection algorithm; simply stated, if the increased degrees of freedom produced by adding an input to a linear regression model outweigh the improvement in that model's performance, the selection process is terminated. Regardless which selection criterion is used by an IVS method, over- or under-selection of inputs, as determined by the termination criterion, may be detrimental to the overall performance of the model.

3.1.3 Objectives

The overall objective of this research is to further develop, refine and compare the four IVS methods described above: PC, PMI, IO, and CNPS, so as to provide better tools for the development of ANN models used for flow forecasting. First, we propose two novel advancements of the IO and CNPS methods: the new IO method builds on previous IO methods and is adapted to quantitatively identify non-useful inputs; the CNPS builds on work by Duncan (2014) and is improved by measuring input usefulness based solely on consistency and eliminating the requirement of an arbitrary threshold for the selection criterion. We compare these two model-based methods to two commonly used model-free methods (PC and PMI). Such a comparison is necessary to demonstrate the advantages and limitations of each approach, and a comprehensive comparison of performance efficacy has not been published in literature before. These methods were chosen based on their distinctive characteristics, which allow for further comparative analysis between IVS methods. Other than comparing model-free and model-based IVS approaches, the selection of the aforementioned IVS methods also allows for the comparison of linear (PC) and non-linear (PMI, CNPS, IO) methods, and low computational effort (PC, CNPS, IO) versus high computational effort (PMI) methods.

Secondly, the impact that standard and predetermined termination criteria, which are listed for each IVS method in Table 3-2, have on ANN model performance is quantified. The model performance with termination criteria-based selected inputs is compared to model performance using a predefined number of inputs determined by each IVS method. This analysis will help answer questions related to optimizing model complexity (measured as the number of inputs used) and model performance. Also, the ANN model performance using IVS is compared to the performance of using a full set of candidate inputs, which provides a baseline, in an effort to quantify the improvement in model performance attributable to each IVS method.

Often, IVS methods are evaluated based on synthetic datasets or a single environmental system, limiting broader use of the developed methods. This research is conducted for two watersheds: the Bow River watershed in Calgary, Alberta and the Don River watershed in Toronto, Ontario (both in Canada). The watersheds have distinct dominant hydrological processes: the first is a relatively large, snowmelt dominated watershed with strong seasonality; the second is a small, flashy, urbanized watershed. Additionally, the Bow watershed has a relatively large dataset available, whereas the Don watershed has a short timeseries – allowing for a comparison of the methods under different data availability conditions. The validation of IVS methods on real-world datasets for two distinct watersheds demonstrates the broader applicability of these methods for future use.

The expected outcomes of improving input variable selection methods is a reduction in data requirements, uncertainty of predictions, and computing power required for these models, and thus improve flow forecasting.

The following section provides details of the structure of the ANN used, the mathematical development of each IVS method, and model performance metrics used for the comparison.

3.2 Methods

This section provides a detailed description of the ANN model configuration, model framework, and IVS methods. The methods described below, and throughout the following chapter, were implemented using MATLAB 2019a. The pseudocode for each IVS method is included in appendix A-1.

3.2.1 ANN model configuration

A generalized process flow diagram that describes the model framework is provided in Figure 3-2. A feedforward multi-layer perceptron ANN is used for this research, which is the most widely used type of ANN flood forecasting (Maier et al. 2010, Abrahart et al. 2012). Multi-layer perceptron ANNs consist of an input, hidden, and output layer, where every node in each layer is connected to every node in the next layer. The ANN parameters (weights and biases) are calibrated using one of many learning algorithms. Using a widely used model type facilitates comparison with existing studies. Moreover, both model-based IVS methods are designed based on the architecture of feedforward ANNs.

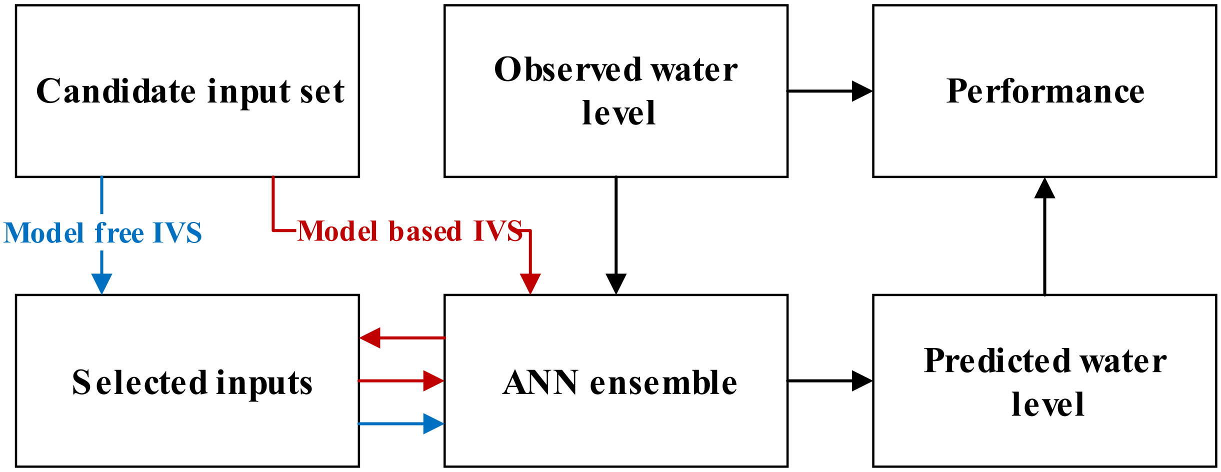

Figure 3-2: A generalized modelling process flow diagram for the model-free and model-based IVS methods; note that for all cases a feedforward multi-layer perceptron ANN model was used.

Each model predicts a single output: either hourly (1, 3, and 6-hour for the Don) or daily (1, 2, and 3-day for the Bow) water levels. A single output node is favoured for ANN models, since multiple models, each with single output variables, will outperform a single multi-output model (Masters 1993, Kaastra and Boyd 1996). For instance, forecasts of 1-day, 2-days, and 3-days are predicted using distinct models, rather than a single model with three outputs.

All models use a single hidden layer with 25 nodes, which is sufficiently sized such that it does not restrict the performance of models with a large input set. While various empirical 'rule of thumb' methods exist for choosing the number of hidden nodes, there is high variability between these methods (Maier and Dandy 2000). While the hidden layer could be optimized based on a systematic trial-and-error approach for each unique model configuration, it is computationally demanding and beyond the scope of this research (Khan et al. 2018). The second-order Levenberg-Marquardt backpropagation algorithm (LMBP) is used to train the neural network. While the first-order backpropagation algorithm is the most widely used algorithm for flood forecasting ANNs, the Levenberg-Marquardt is also commonly used and is considered more efficient than the simpler backpropagation algorithm (Maier et al. 2010, Abrahart et al. 2012).

Typically, for ANNs, datasets are partitioned into three subsets: training, validation, and testing. The training subset is used to calibrate the ANN’s weights and biases, the validation subset is used to terminate training (and prevent overfitting), and the testing subset is used to evaluate the model performance. In this research, the dataset is partitioned into training-validation-testing blocks of 60%-20%-20%, respectively. K-Fold Cross-Validation (KFCV) is used to train a collection of networks, which is illustrated in Table 3-1. Using a cross-validation method such as KFCV reduces the chance of the validation partition biasing the results, reduces the uncertainty in the predictions, and is useful for smaller datasets. For both catchments in this research, 5 folds (4 for training and validation, and 1 independent fold for testing) are used, that are 2-years and 2-months for the Bow and the Don Rivers, respectively.

Table 3-1: An illustrative example of 4-Fold Cross-Validation dataset partitioning for training ANN models with three data partitions: training, validation and testing.

	Fold				
Iteration	**1**	**2**	**3**	**4**	**5**
1	Validation	Training	Training	Training	Testing
2	Training	Validation	Training	Training	Testing
3	Training	Training	Validation	Training	Testing
4	Training	Training	Training	Validation	Testing
⋮	⋮	⋮	⋮	⋮	⋮
1000	Training	Training	Training	Validation	Testing

KFCV is used to capture the uncertainty associated with the selection of the calibration (which includes the training and validation) partitions, then each unique fold configuration is trained 250 times to capture the uncertainty associated with the ANN initialization, which assigns random values to weights and biases to begin the LMBP training algorithm. This multi-start initialization approach helps reduce overfitting issues and uncertainty with ANN predictions. Collectively, the four unique calibration configurations and 250 multi-start initialization results in 1000 ANN models, which are referred to as an ensemble; discrete models are referred to as ensemble members.

3.2.2 Input variable selection

The four IVS methods evaluated in this paper are summarized in Table 3-2, along with the selection and termination criteria used for each method. These criteria are described in detail in the following sections. All the IVS methods are applied using the calibration partition and the ANN performance is measured using the independent test partition. Pseudocode for each IVS algorithm is included in Appendix A, which provides additional description for each algorithm and facilitates implementation.

Each IVS method is used to reduce the number of inputs based on a pre-established termination criterion (listed in Table 3-2), which is common practice for IVS algorithms. However, using an IVS termination criterion may result in the over- or under-selection of input variables. For example, if PC selects ten inputs and PMI selects five inputs, assuming there are ten reasonably useful inputs, the PC-based ANN may perform better than the PMI-based model, since more useful data is included. The lesser performance is not reflective of PMI as a poor selection criterion, rather that the final selection is repressed by a too strict termination criterion; the five inputs selected by PMI may all be useful however the selection was terminated too early. This highlights the significance and impact of the termination criteria used for IVS.

In order to address the risk of non-optimum termination criteria, in addition to selecting inputs using a termination criteria, each IVS method is used to select the most relevant inputs (which we have selected as 10% and 20% of the candidate set for each river) without regard for any termination criteria. Such will showcase the capability of each IVS method to identify the most useful inputs and allows for a direct comparison between input reduced models that have the same level of complexity (i.e., the same number of selected input parameters).

Note that for methods where inputs are not selected iteratively (i.e., CNPS and IO), and thus the inputs are not ranked based on importance, the selection criteria to rank inputs are described in Table 3-2 and in the appropriate subsections in the methods section; they are used to select a predefined number of inputs.

Table 3-2: A summary of the four IVS methods used, including the selection and stopping criteria used for this research.

IVS Method	Selection type	Stepwise method	Selection criteria	Stopping criteria	Computational effort
Partial Correlation	Model-free	Forward selection	Max PC	AIC increase	Low
Partial Mutual Information	Model-free	Forward selection	Max PMI	AIC increase	High
Input Omission	Model-based	Global selection	Max AIC	p-value ≤ 0.01	Medium
Combined Neural Pathway Strength	Model-based	Global selection	Max α[1]	α ≥ 0.95	Low

[1] Instances with more than 1 input with α = 1 are ranked using the method described in section 3.2.2.4

3.2.2.1 Partial correlation

This method uses a forward selection algorithm to select useful inputs from a candidate set C based on the maximum PC between a given candidate input C_j and the output Y, controlling for the effects of the set of inputs that have already been selected, S. Note while the selected set S is typically referred to as a subset of the candidate set C, in this algorithm the number of candidate inputs in C is reduced at each selection, as selected inputs are removed from the set of candidates. For the first selection, S is an empty set and the first input is selected as having the maximum squared linear correlation between the candidate C_j and the output Y. Note that both the PC and

PMI-based input selection methods were adopted directly from existing studies (Sharma 2000, Bowden et al. 2005b, May et al. 2008b, He et al. 2011). However, one distinction in the implementation of PC in this research that distinguishes it from previous studies is the use of R^2 instead of R to avoid inputs with a strong negative correlation being selected last. For a set of selected inputs S and a set of J candidate inputs, for candidates $j = 1{:}\,J$, the partial correlation between C_j and output Y, controlling S is given by:

$$\mathrm{PC}(C_j; Y|S) = R(v; u)^2 = \left(\frac{\mathrm{cov}(u, v)}{\sigma_u \sigma_v}\right)^2 \tag{1}$$

where u and v are the residuals, given by the following equations:

$$u = C_j - \widehat{C_j}(S) \tag{2}$$

$$v = Y - \widehat{Y}(S) \tag{3}$$

where $\widehat{C_j}$ and $\widehat{Y}$ are the least square estimates of C_j and Y, respectively.

The algorithm for PC first calculates the correlation for each input and selects the input having the highest correlation. Next, the residual u is calculated for the selected input(s), using equation (2), and the residual v is calculated for each remaining candidate input, using equation (3). The input corresponding to the highest partial correlation, which is described in equation (1) for PC, and is a function of u and v, is moved from the candidate set to the selected set. At each step, the AIC is calculated, as per the following equation:

$$\mathrm{AIC}_j = n \log_e\left(\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2\right) + 2p \tag{4}$$

where n is the number of samples and p is the number of parameters in the linear regression model.

This selection process is repeated until the AIC increases if using the termination criterion, the desired number of inputs is reached (3, 6, 12, etc. for the non-termination criteria experiments), or no inputs remain in the candidate set. The pseudocode for PC is provided in the appendices.

3.2.2.2 Partial mutual information

This selection method is conceptually similar to the PC-based method described above but uses mutual information (MI) rather than linear correlation for input selection. The PMI value is the shared entropy between output Y and candidate input $\mathrm{C_j}$ that is not also already contained in S, expressed as $\mathrm{MI(C_j; Y|S)}$. The PMI-based approach uses the same selection algorithm as for PC, however the linear estimators $\hat{\mathrm{C}}(S)$ and $\hat{\mathrm{Y}}(S)$ are replaced with non-parametric estimators $\mathrm{E}\big[\mathrm{c_j|S = s}\big]$ and $\mathrm{E[y|S = s]}$, respectively, and the selection criteria of partial correlation, PC, is replaced with the partial mutual information PMI. The non-parametric kernel regression estimation (i.e., the conditional expectation of x, given S) is given by the expression:

$$\mathrm{E[x|S = s]} = \frac{1}{\mathrm{n}} \frac{\sum_{\mathrm{i=n}}^{\mathrm{n}} \mathrm{x_i K_h(s - s_i)}}{\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{K_h(s - s_i)}} \tag{5}$$

where n is the total number of sample observations, S is the selected set, x is either $\mathrm{c_j}$ or y, and $\mathrm{K_h}$ is the Gaussian kernel function, given by:

$$\mathrm{K_h(x - x_i)} = \frac{1}{\left(\sqrt{2\pi \mathrm{h}}\right)^{\mathrm{d}} \sqrt{|\sigma|}} \exp\left(-\frac{\mathrm{(x - x_i)^T \sigma^{-1} (x - x_i)}}{2\mathrm{h}^2}\right) \tag{6}$$

where d is the dimensionality of x, σ is the sample covariance matrix, and h is the kernel bandwidth given by:

$$h = \left(\frac{4}{d+2}\right)^{\frac{1}{d+4}} n^{\frac{-1}{d+4}} \tag{7}$$

The PMI is then calculated as follows:

$$\mathrm{PMI}(C_j; Y|S) = \mathrm{MI}(u; v) \approx \frac{1}{n}\sum_{i=1}^{n} \log_e \left[\frac{f(u,v)}{f(u)f(v)}\right] \tag{8}$$

where u and v are calculated using equations (2) and (3) using the respective non-parametric kernel estimators $\widehat{C_j}(S) = E\left[c_j|S = s\right]$ and $\widehat{Y}(S) = E[y|S = s]$, described in equation (5). $f(u)$, $f(v)$, and $f(u, v)$ are probability density functions of u, v, and joint u and v, estimated using the following generalized expression:

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) \tag{9}$$

where n is the sample size and K_h is the kernel function, provided in equation (10).

The AIC, which is used as the termination criteria, is calculated using equation (4) where the number of parameters p is replaced with the effective degrees of freedom ν. For non-parametric kernel regression, the effective degrees of freedom is given by the trace of the hat matrix H:

$$\nu = \mathrm{tr}(H) \tag{11}$$

where the hat matrix H is given by:

$$H_i(x) = \frac{K_h(x_i - x)}{\sum_{j=1}^{n} K_h\left(x_j - x\right)} \tag{12}$$

3.2.2.3 Input omission

The method proposed in this research for IO estimates input usefulness based on the change in performance following input omission without retraining, adapting and improving previous methods used that are described in existing studies (c.f. Abrahart et al., 2001; Setiono and Liu, 1997). The complexity-based performance measure, AIC, is used as the performance criteria in this method, as it permits for a small increase in error to be negated due to a slightly lower model complexity. The AIC is calculated using equation (4) where p is the number of ANN parameters (the total count of the weights and biases) and $\widehat{\mathrm{Y}}$ is given by:

$$\widehat{\mathrm{Y}}_{\mathrm{j}} = \mathrm{net}_{\mathrm{j}}\big(\mathrm{C}|\ \mathrm{C}_{\mathrm{j}} \notin \mathrm{C}\big) \qquad (13)$$

where $\widehat{\mathrm{Y}}_{\mathrm{j}}$ is the output of $\mathrm{net}_{\mathrm{j}}\big(\mathrm{C}|\ \mathrm{C}_{\mathrm{j}} \notin \mathrm{C}\big)$, which is the ANN trained on C with input C_{j} omitted.

Inputs are selected based on the p-value corresponding to the null hypothesis that $\mathrm{AIC}_{\mathrm{j}}$ and $\mathrm{AIC}_{\mathrm{C}}$ are samples from the same population against the alternative hypothesis that $\mathrm{AIC}_{\mathrm{j}}$ is larger than $\mathrm{AIC}_{\mathrm{C}}$ using a two-sample Kolmogorov-Smirnov (KS) test. This research uses a p-value of 0.01 for the null hypothesis test. The KS test cannot be used to rank input parameters, therefore if IO is being used to select a predefined number of inputs, inputs are ranked based the median $\mathrm{AIC}_{\mathrm{j}}$ values within the ensemble, such that the omitted input corresponding to the largest increase in median AIC is the most useful.

3.2.2.4 Combined neural pathway strength

The CNPS method used in this research is a modified, improved and more generalised version of the method proposed by Duncan (2014), where the neural pathway strengths are calculated for an ensemble of models, which are then used to characterize inputs as ‘excitatory’ and ‘inhibitory’. In

Duncan's method, the CNPS values are calculated as the matrix multiplication of the input-hidden and hidden-output weight matrices:

$$\mathrm{CNPS}_{ik} = \mathrm{W}_{ij} \cdot \mathrm{W}_{jk} \tag{14}$$

where W_{ij} and W_{jk} are weight matrices with dimensions $i, j, k,$ which correspond to the number of input, hidden, and output nodes respectively. In Duncan's original method, inputs are selected based on the ensemble interquartile range (EQR) metric, which is defined as follows:

$$\mathrm{EQR_j} = \frac{\min(|\mathrm{Q}_1|, |\mathrm{Q}_3|)}{\max(|\mathrm{Q}_1\ |, |\mathrm{Q}_3|)} \cdot \mathrm{sgn}(\mathrm{Q}_1) \cdot \mathrm{sgn}(\mathrm{Q}_3) \tag{15}$$

where Q_1 and Q_3 are the 25th and 75th percentiles of the $\mathrm{CNPS_j}$ distributions for input j. EQR values are within the range of [-1,1] and values greater than 0 are considered satisfactorily excitatory or inhibitory. However, while using this approach, it was found that using an EQR of 0 as a threshold for IVS resulted in over-selection. Consequently, the EQR threshold was incrementally increased to an arbitrary value in order to select an optimum number of inputs based on model performance. Choosing an arbitrary non-zero EQR value is undesirable, since it required a trial-and-error approach and has no statistical meaning. Moreover, the rate at which inputs are exhibiting excitatory or inhibitory behaviour is not known. Thus, in this research, the EQR metric was abandoned, in favour of a statistical threshold. Inputs are ranked based on the percentile of dominant excitatory or inhibitory behaviour, inferred by the signs of the CNPS values. For example, if a threshold of 95% is selected, it ensures that for each selected input, inputs are exhibiting either excitatory or inhibitory behaviour at least 95% of the time within the ensemble. The proposed alternative selection criterion for CNPS was found to produce an input ranking very similar to the original EQR metric. The main advantage of the proposed selection criterion is that input score is based solely on the consistency of the input behaviour (i.e., excitatory or inhibitory), unlike the previous (existing) EQR method. The previous method indicates usefulness or non-usefulness based on the sign of the EQR measure, which is positive only if at least 75% of the inputs are behaving consistently. The threshold of 75% was found to be much too lenient and reducing the number of inputs requires an exhaustive trial-and-error approach. The proposed method allows for the number of inputs to be constrained by changing the threshold of consistent behaviour (e.g., if 95% is used, all of the selected inputs exhibit excitatory or inhibitory in at least 95% of the models in the ensemble). The behavioural consistency of CNPS values, using the proposed criterion (i.e., expressed as a percentage), α, for each input j is determined as follows:

$$\alpha_j = \frac{\max\left(\sum(\mathrm{CNPS_j} > 0), \sum(\mathrm{CNPS_j} < 0)\right)}{n} \tag{16}$$

where $\mathrm{CNPS_j}$ is the distribution of CNPS scores and n is the number of ensemble members. Since multiple inputs may be identified at the significance value of $\alpha_j = 1$; these inputs can be ranked based on the relative range of their CNPS values:

$$\mathrm{score} = \frac{\min(\mathrm{CNPS_j})}{\max(\mathrm{CNPS_j})} \tag{17}$$

Inputs with a smaller range between the minimum and maximum CNPS values will receive a higher score, with a theoretical optimum score of 1, which indicates that the input exhibits exactly the same behaviour across the entire ensemble.

3.3 Results and discussion

Best practices in hydrological modelling suggest that model performance be assessed based on multiple performance measures, as different measures capture different model characteristics (Maier et al. 2010, Ewen 2011). There is a wide variety of commonly used measures; this research utilises four common measures, including Root Mean Squared Error (RMSE), Nash-Sutcliffe Efficiency (NSE), Mean Absolute Error (MAE) and the Persistence Index (PI). The PI is similar to the NSE, but specialized for forecasting models; instead of the normalizing squared residuals being based on the observed value, PI uses the observed value lagged by the model's lead time (Kitanidis and Bras 1980, De Vos and Rientjes 2005, Abrahart et al. 2008). These error measures are chosen from three distinct error measure taxonomies used for neural network assessment, being based on square residuals (RMSE and NSE), absolute residuals (MAE), and timing (PI) (Maier et al. 2010). While NSE and RMSE are essentially the same, they are both included in this paper due to their widespread use in previous hydrological studies (Ewen 2011).

The ANN models for both watershed systems were trained with the complete candidate sets, and IVS-reduced inputs (both using the termination criteria and predefined numbers of inputs) to predict the stage at the target station. A comparison of model performance using the four metrics, RMSE, NSE, MAE and PI, are summarised in Figure 3-3 (for the Bow River) and Figure 3-4 (for the Don River) for the 1-day and 1-hour lead times. Note that these metrics were calculated only using the independent testing dataset (the training and validation dataset are excluded from all performance evaluation). In addition, the results are calculated for all the models within each ensemble (i.e., including the cross-validation and multi-start scenarios) so as to quantify the uncertainty of the predictions. Thus, the figures show the 25th and 75th percentile values (blue boxes), the median value (red line), and outliers (red crosses) of the predictions rather than deterministic results. Additional results for the other lead times for both rivers are included in appendix A-2.

These figures show the performance of the flood forecast models using different combination of inputs (on the horizontal axes) and each performance metric on the vertical axes. On the horizontal axis of each subplot, the left most values are from the base model (which includes all the candidate inputs), followed by each of the four IVS methods, first using the predetermined termination criteria, then the two cases with predefined numbers of inputs. The two cases for the predefined number of inputs are 10% of the candidate inputs: 3 for the Bow River and 6 for the Don River; and 20% of the candidate inputs: 6 for the Bow River and 12 for the Don River. For example, in Figure 3-3 the PC results are indicated by PC (17), PC (3), and PC (6), which corresponds to 17 inputs determined by the termination criteria, followed by 10% (3 inputs) and 20% (6 inputs) of the 30 candidate inputs.

Figure 3-3 and Figure 3-4 demonstrate the effectiveness of using an ANN approach for flood forecasting: the error metrics indicate high performance for each metric. The RMSE and MAE are low (approximately 1% relative to of the observed values) and the NSE is high (with median values about 0.9 for the Bow, and above 0.85 for the Don). Similarly, the median PI values are positive indicating that the predicted values are an improvement over the last known flow values. Similar results can be seen for models with higher lead times (results in Appendix B), and as expected, the overall performance decreases and the variance increases, as the lead time increases.

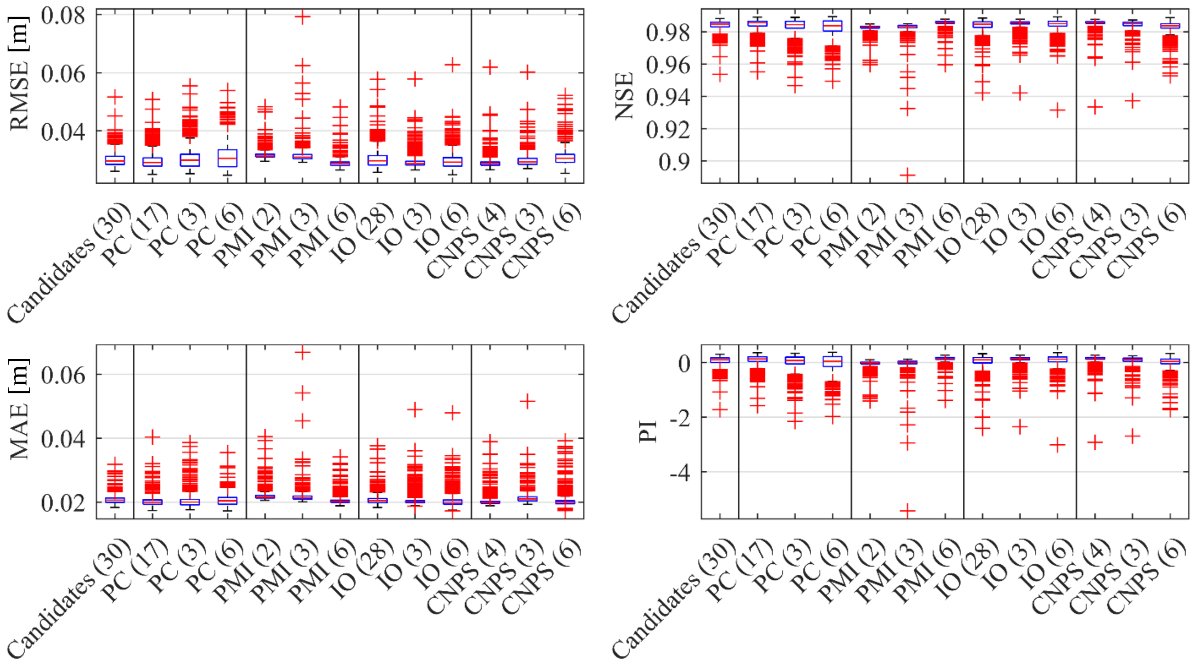

Figure 3-3: Comparison of ANN model performance (RMSE, NSE, MAE, and PI) for the Bow River for the 1-day lead time for models that use all candidate inputs (30), termination criteria-based inputs (variable), 10% of all inputs (3), and 20% of all inputs (6) for each of the 4 IVS methods.

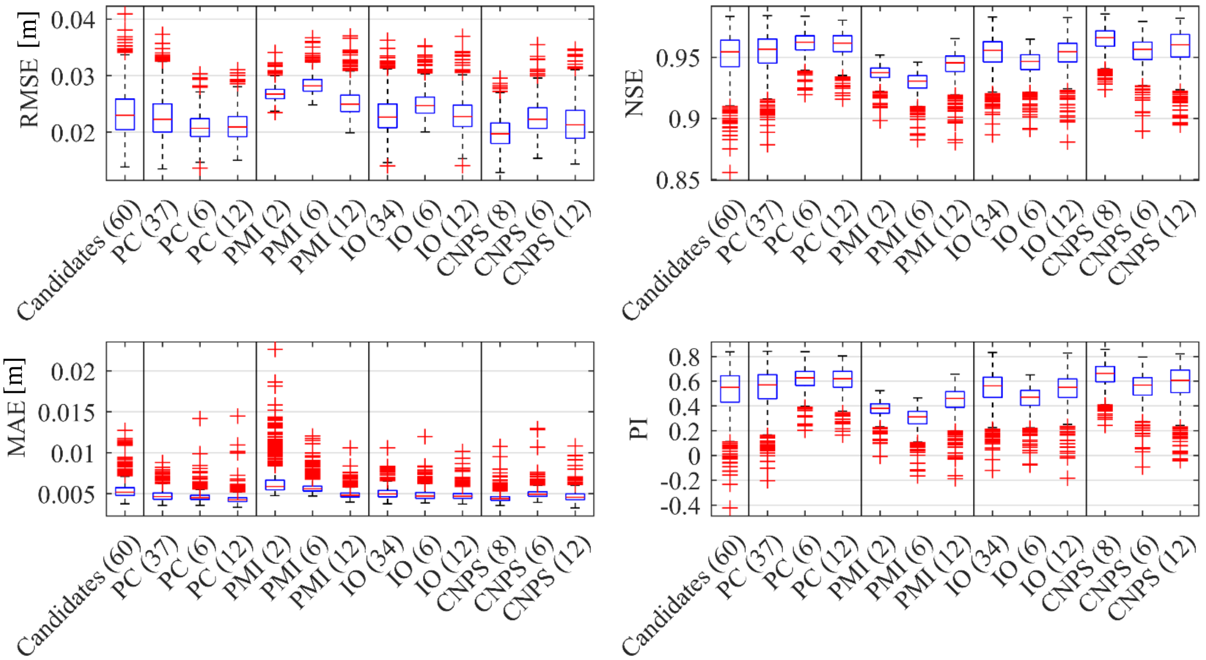

Figure 3-4: Comparison of ANN model performance (RMSE, NSE, MAE, and PI) for the Don River for the 1-hour lead time for models that use all candidate inputs (30), termination criteria-based inputs (variable), 10% of all inputs (6), and 20% of all inputs (12) for each of the 4 IVS methods.

Overall, this suggests that it is possible to achieve high performance of flow forecasting models using an ANN approach. However, an analysis of the impact of selecting inputs (based on different IVS methods) can help further refine these models (in terms of performance, computational efficiency, etc.), which are detailed in the following subsections, along with detailed information on which inputs were selected for each case.

3.3.1 Model performance with all candidate inputs

The correlation and timeseries plots for the base models (i.e., those that use all candidate inputs) with one timestep leads (1-day for the Bow, and 1-hour for the Don River) are provided in and Figure 3-6, respectively. which provide a general sense for the model performance. Both figures include 99th percentile bars, which indicate the amount of uncertainty within the model ensembles, which is owed to the multi-start and KFCV method employed. shows that the results of the independent testing dataset are very similar to the calibration (training and validation) datasets. Based on these figures, both models exhibit slight under-prediction of high water levels (hence flowrates) of up to 0.5m from the ensemble median in several cases and approximately 0.1m from the confidence interval for the poorest predicted timestep. Underprediction is more pronounced for the Don River, which is the more urbanized and flashier watershed. From Figure 3-6, it is apparent that both model outputs occasionally exhibit visible timing error during peak events; however, for the Bow River, this difference is generally captured within the 99th percentile uncertainty bands. This timing error, which is indicated by the predicted water levels being shifted to the right of the observed levels, is the primary focus of the following chapter.

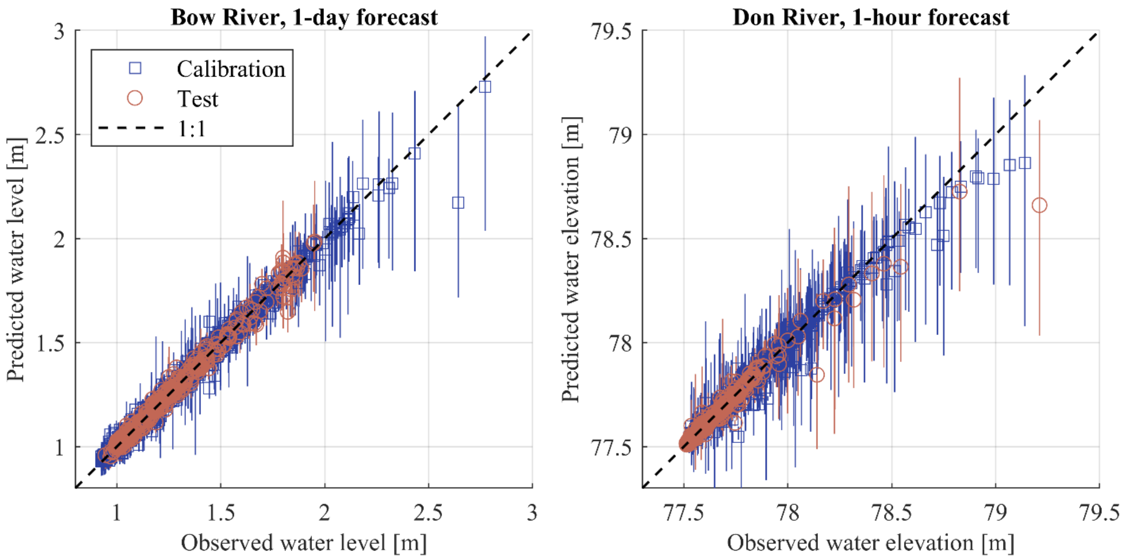

Figure 3-5: Correlation plots showing the observed and predicted (using all candidate inputs) water levels for the Bow and the Don Rivers for both the calibration (training and validation) and testing datasets; note that the predicted values include the model ensemble median (symbols) and 99th percentile range (lines).

While a direct comparison with other hydrological systems offers little insight, as different systems have unique physical characteristics and data availability, it can provide a broad sense of whether the models are exhibiting reasonable performance. Out of the four error measures, RMSE and MAE are not normalized, and while useful for comparing between the various IVS configurations, they are not useful for comparison between different systems or external models (Mosavi et al. 2018). The NSE values for both rivers are considered acceptable (above 0.8) for five of six lead times evaluated (Mosavi et al. 2018). The 6-hour Don River forecast model demonstrated a poor NSE of approximately 0.5 as indicated in the Appendices (section A-2); however, based on discussion with the local watershed authority and the approximate lag times between upstream and downstream stations (inferred by the cross-correlation analysis used to select the initial candidate set), poor forecasting performance is expected for a lead time of 6-hours.

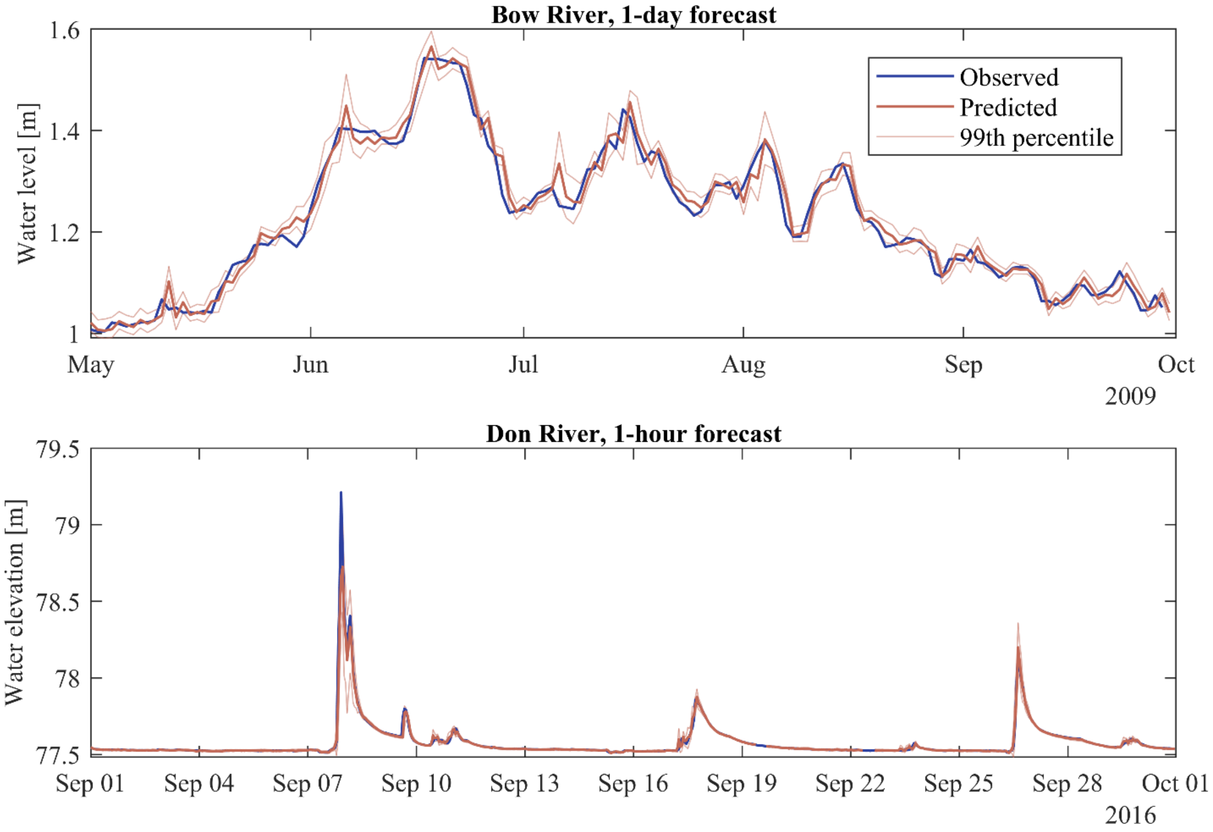

Figure 3-6: Timeseries plots of observed and predicted (using all candidate inputs) water levels for the Bow and the Don Rivers, for a section of the test dataset; note that the predicted values include the model ensemble median and 99th percentile range.

Next, while the Bow River demonstrated a strong NSE performance across all the lead times considered (1, 2 and 3-day), this strong performance is somewhat superficial. The Bow River has high seasonality, which can produce a model with a misleading NSE value (Meshram et al. 2018). Thus, the PI is a better indicator of forecasting strength of the model, which decreases as the lead time increases, as is expected for forecast accuracy, unlike the NSE, which remains high across all lead times. In contrast to the Bow River, the Don River has no significant seasonality, as discussed in section 2.1, therefore the NSE is a suitable performance measure as the explained fluctuations from the mean water level are due to hydrological events and are not overshadowed by seasonality within the system. The performance of these base models is improved using IVS methods (as illustrated in Figure 3-3 and Figure 3-4), and discussed in detail in the following section. As discussed in section 3.1.2, reducing the number of model inputs can improve convergence during training, and lower the complexity and data requirements of models.

3.3.2 Model performance with IVS

The majority of IVS models exhibit stronger performance compared to the non-IVS models. Generally, the median performance improves slightly and range (defined as the difference between the minimum and maximum performance within the ensemble) decreases. This can be attributed to the fact that the IVS models have fewer non-useful inputs (see Figure 3-3 and Figure 3-4). The IVS models exhibit a decrease in performance when too few inputs are selected; such is the case for the PMI models where the termination criterion produces a model with only two inputs for the both the Bow and the Don, both of which perform poorly. Similarly, if too many inputs are selected (e.g., the IO models with the default termination criteria), the model performance does not improve compared to the base models.

The selection order for the models with the termination criteria and models that select 20% of the candidate set is provided

Table 3-3 and Table 3-4, where the input selection order is indicated numerically as well as by the colour intensity of the table cells. The selection and termination criteria for each method are shown graphically in Figure A-1 and Figure A-2 in the Appendices, for the Don and the Bow Rivers, respectively. These figures show the relative change in the magnitude of the selection criteria for the selection of the first 20% of the highest ranked inputs; note that the termination criteria may stop the selection before or after the number of inputs included in these figures. A detailed analysis and comparison of using termination criteria versus using a predefined number of inputs is provided in the following section.

3.3.2.1 IVS with termination

In these models, the number of inputs is determined by a predefined termination criterion that terminates input selection when met, which is outlined Table 3-2. For both the Bow and the Don cases, PC selects roughly half of the available candidate inputs. Expectedly, the performance of the model ensembles is similar to that of the base models given the large number of inputs selected from the candidate set. For the next IVS case, PMI selects only 2 to 3 inputs in all cases. Consequentially, the performance of these models is very poor, which suggests poor or under-selection of inputs. This early stopping is partly a result of the high estimated degrees of freedom of the non-parametric kernel regression models, which produces a local minimum AIC after only two to three inputs are selected. The relative number of inputs determined by these methods broadly agree with May et al. (2008b), who observed that the PC termination criterion is more lenient than the PMI-based counterpart.

For the model-based IVS methods, IO for Bow River selected nearly all the candidate inputs, whereas for the Don, roughly half. While developing the IO method used in this paper, the use of a KS test was appealing due to its suitability for non-parametric distributions. This criterion produced a reasonable selection: roughly half of candidate inputs for the Don River. However, while validating the method on the Bow River models, it became clear that the criterion was not stringent enough, resulting in an over-selection. Nonetheless, the IO method is reasonably capable of identifying useful inputs, which is evidenced by the selection order illustrated in Tables 5 and 6. The selection order for IO in these tables indicates that non-useful inputs, such as temperature which is not selected in the best performing models, are among the last selected by the IO method, while the first selections agree with those of IVS methods that produce strong performing models. This claim is further supported by the performance of IO models that have used a predefined number of inputs, which is discussed in the following section. Lastly, for the CNPS method: in both cases, this method selected between 10-20%, and performed similarly to the fixed input cases, suggesting that the termination criterion for CNPS performs well and consistently.

3.3.2.2 IVS for a predefined number of inputs

In this section, the optimality of the number of inputs determined using predefined termination criteria is assessed by comparing these inputs with two cases of models that have a predefined number of inputs (selected as 10% and 20% of the candidate input set for each watershed). The use of predefined number of inputs means that each IVS method can be compared against each other for the same number of inputs. This approach allows for an assessment of each IVS methods performance whilst controlling for the number of inputs (or level of complexity) of the ANN and the termination criteria used, thus indicating which method selects the best inputs for improved model performance for the same level of model complexity.

For a predefined number of inputs for the Bow model, PC demonstrates poor performance compared to the termination criterion counterpart, and models for other IVS methods that have the same number of inputs, indicating that PC does not accurately rank the most useful inputs within the candidate set and therefore isn’t a good IVS method. Specifically, using

Table 3-3, PC failed to select daily maximum inputs for station BH001, which are included in the selection of better performing models. In the case of the Don, the performance of PC improved which suggests that the termination criterion was too lenient. Next, PMI overwhelmingly favoured autoregressive inputs, selecting exclusively autoregressive inputs for the Bow and all the autoregressive inputs for the Don. In the case of the Don, PMI first selected all 6 inputs for the target station, followed by all 6 inputs for a nearby upstream station (HY022).

Table 3-3: Bow River IVS ranks with termination (upper) and for the first 6 (20% of candidate set) inputs (lower).

Parameter:	WL									WL									Precipitation			Temperature								
Station ID:	BH004									BB001									3031093			3031093								
Daily	Mean			Min			Max			Mean			Min			Max			Sum			Mean			Min			Max		
Lag (days)	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1	3	2	1
Termination																														
PC	5	2	1			10		8	7	14		12	9		11	15	4	3	16		17					13				6
PMI			1						2																					
IO	15	3	1	16	19	13	17	5	2	7	10	9	6	12	11	14	8	4			26	22	21	20	23	25	24	28	27	18
CNPS		2	1					3	4																					
Fixed																														
PC	5	2	1														4	3												6
PMI		4	1		5	3		6	2																					
IO		3	1					5	2				6					4												
CNPS		2	1	6				3	4									5												

Table 3-4: Don River IVS ranks with termination (upper) and for the first 12 (20% of candidate set) inputs (lower).

Parameter:	WL						WL						WL						WL						WL						Precipitation												Precipitation												Temperature					
Station ID:	HY019						HY017						HY093						HY080						HY022						HY008												HY027												31688					
Lag (hours)	6	5	4	3	2	1	6	5	4	3	2	1	6	5	4	3	2	1	6	5	4	3	2	1	6	5	4	3	2	1	12	11	10	9	8	7	6	5	4	3	2	1	12	11	10	9	8	7	6	5	4	3	2	1	6	5	4	3	2	1
Termination																																																												
PC		30		13	11	1	26	36	19	18		20							37	15	21	3	10	2	25	34	6	12	7	5	35		31	29			22	27	33	9	4	17		32	28		24		14	8	23	16								
PMI					2	1																																																						
IO	34		24	7	2	1	12	10	21	16	9	14	29	13	15	11	19	28		30			25	3		17	5	4	6	20										23	8											31			33	27	26	22	18	32
CNPS					8	1																		2			6	3	4												5										7									
Fixed																																																											
PC					11	1																3	10	2			6	12	7	5										9	4									8										
PMI	6	5	4	3	2	1																			12	11	10	9	8	7																														
IO				7	2	1	12	10			9					11								3			5	4	6												8																			
CNPS				11	8	1								10										2			6	3	4												5	12									7	9								

These models generally perform better than the termination criterion models, which implies that the criterion is too strict. However, the shortcomings of PMI are best evidenced by the PI measure, which is much lower than the model with all candidates in Figure 3-3 and Figure 3-4. One cause might be that the PMI selection method favours the target station inputs, which may be less useful for forecasting compared to upstream, exogenous inputs. The other IVS models, which utilize more upstream flow stations, have better PI performance.

The models for IO demonstrate improved performance while using fewer inputs, indicating that this method is reasonably capable of ranking inputs, however, the termination criterion used in this research is much too lenient (given the large number of inputs selected for each watershed). Finally, CNPS models typically exhibit slightly poorer performance while using a predefined number of inputs, even while the number of inputs determined using a termination criterion is in-between the two cases that use a predefined number of inputs, suggesting that the termination criterion determined nearest to the optimum number of inputs. This underscores the utility and effectiveness of the proposed statistical threshold proposed and implemented in this research for the CNPS method.

3.3.3 Discussion

Ultimately, using a termination criterion consistently produces better performing models than using a predetermined number of inputs. This is true across both the Bow and Don watersheds and for the different IVS methods. For example, in Figure 3-3, IO (3) outperforms IO (28), which over-selects inputs using the termination criterion. However, the over-selection is not observed at larger lead times, where using a termination criterion produces slightly better performing models than IO (3) and IO (6). Next, in Figure 3-3, the termination criteria for PC (17) over-selects while the PMI

(2) models under-select, since they both exhibit poorer performance compared to their counterpart selections with 12 inputs. Resultantly, it is recommended that the number of inputs be determined using a systematic approach, where each input selection is informed by an IVS-based input ranking. This approach may use any suitable performance measure(s), or a complexity-based measure for determining the optimum number of inputs. For example, a method such as CNPS may be used to rank input usefulness. The input rank may be used inform a forward selection scheme where inputs are incrementally added to a neural network, until the performance ceases to increase or the increased model complexity is unwarranted. All four IVS methods reviewed in this study could be used to rank input usefulness and are suitable for such an approach.

All four IVS methods identify the most recent autoregressive input (daily mean) as the most useful. Often one or two more autoregressive inputs are selected before methods select exogenous inputs. In fact, the best preforming models for the Bow are achieved using only autoregressive inputs. Next, three of the four IVS methods use one or more of the autoregressive daily maximum and upstream daily maximum data. Precipitation and temperature are seldom selected, unless the IVS method's termination criteria is very lenient, as is the case for PC and IO. The preference of autoregressive inputs in all models results in time-delayed predictions, which is evidenced by the strong NSE but poor PI, which is most apparent for longer lead times (see Appendix B). The preference of autoregressive inputs for model-free IVS methods is likely due to the strong dissimilarity between autoregressive and exogenous signals, whereas for model-based methods, it may be attributed to the objective function (MSE) of the ANN. The utilization of MSE for ANN training has been associated with time-delayed predictions, subsequently, model-based IVS methods will favour autoregressive inputs. As such, model-based IVS methods cannot reliably avoid time-delay issues. For the Don, similar to the Bow, the first selection is the most recent autoregressive input (see

Table 3-3 and Table 3-4). Next, multiple inputs at upstream gauge HY022 are selected by all four methods, as well as HY080 by PC, IO, and CNPS. These three methods also each utilize one to four precipitation inputs. The best performing models for the Don rely on exogenous inputs, which is evidenced by the difference in performance of the PC, IO, and CNPS selections, and the autoregressive dominated PMI selection, which is discussed in section 3.3.2.2.

Next, this research demonstrated an application of KFCV, which, coupled with multi-start, is used to generate model ensembles. Using KFCV, or a different variation of cross-validation, is strongly encouraged: it eliminates the subjective discretization of the training and validation data and quantifies the associated uncertainty. Such methods are simple to implement, yet often lacking in existing studies of data-driven flood forecasting models. Furthermore, provided advances to computing power, thousands of ANNs can be trained in a short amount of time.

It is worth restating that input usefulness is a not binary but exists on a spectrum. For real-world datasets in which usefulness is measured based on model performance, whether an input is useful corresponds to a point at which its inclusion in the model will not consistently improve the performance of the model. Consequently, the usefulness of an input is a function of both the performance measures chosen for the model and the selection of model hyperparameters.

As demonstrated in this research, improvements to model performance owed to the application of IVS methods should be assessed using multiple distinct measures. Next, some of the uncertainty associated with hyperparameter selection is considered in this research by using an ensemble-based modelling approach, which demonstrates how the uncertainty associated with the selection of two groups of hyperparameters is quantified; these include the calibration dataset partitioning (KFCV) and the initial values for ANN weights and biases (multi-start).

Lastly, it is important to highlight that the base models, which feature all candidate inputs, perform reasonably well. While the performance is marginally improved with the removal of non-useful inputs, the input reduced models are not drastically superior to non-input reduced models in the context of flood forecasting capabilities, which is largely attributable to the strength of the LMBP training algorithm and ANN structure. The benefits of removing non-useful inputs are better demonstrated by the narrower range in ensemble performance, illustrated in Figure 3-3 and Figure 3-4. Next, the input reduced models are much less computationally demanding than those with the full set of candidates. While computational expense is not a constraining factor for simple single hidden layer ANNs, more complex architectures such as Bayesian or Fuzzy ANNs may receive greater benefits. Finally, in practical applications, there may be value in having redundant inputs, such as in the event of a sensor failure; if an IVS method identifies inputs as non-useful, it does not mean that the monitoring location is not hydrologically relevant, or important with respect to the overall purpose of the model.

3.4 Conclusions and recommendations

This research evaluated four different IVS methods for ANN models for the Bow and the Don Rivers. Two methods were improved: CNPS by developing an alternative selection criterion that places increased emphases on the characterization of input behaviour and IO by describing an ensemble-based quantitative selection criterion, which is lacking from existing work. PC demonstrated reasonable performance for both watersheds, however using a predefined termination criterion causes an over-selection of inputs and the resultant models are outperformed by other IVS methods in most instances. PMI suffered early-stopping, as the termination-based selection was too strict. PMI also favoured autoregressive inputs, which resulted in poor model performance, most notably in the Don River models. It is possible that modifications to PMI such as using a non-Gaussian kernel, or a scaling factor, may yield improvements in input selection. IO demonstrated reasonably strong performance, however the termination criterion used in this research is not recommended for future use, as it was too lenient and inconsistent. IO may be improved by making changes to the method, by exploring topics such as explicitly evaluating the benefits of retraining after omission, omitting more than one input at a time, or using different performance criteria. The modified CNPS demonstrated the strongest and most consistent performance amongst the IVS methods evaluated, which highlights the significance and impact of the proposed improvements to this method.

This research produced results for two reasonably distinct hydrological systems, the Bow and Don Rivers. This type of validation is important for ensuring that methods are sufficiently robust to be useable for different hydrological regions. In this research, the results for both basins were generally in agreement. However, some notable exceptions include the PMI selection for the Bow, which performs adequately as an autoregressive model, however an autoregressive model for the Don exhibits very poor performance. Next, the selection criterion for IO selected a drastically different number of inputs for the Bow and Don, which was not observed using other IVS methods, indicating that the selection criterion proposed in this research may need further refinement before future use. Other methods, such as PC and CNPS, demonstrated reasonable and consistent selections both watersheds.

Next, this research determined that the use of a termination criterion is not consistent in selecting the optimum number of inputs. Instead, it is recommended that IVS be used to rank input usefulness, after which the number of inputs is determined on a case-by-case basis, based on a systematic evaluation of model performance. Future research topics may include the coupled optimization of ANN model inputs and hyperparameters, and further refinements to CNPS-based IVS methods.

This chapter demonstrated that IVS can be used to drastically reduce the number of inputs for flow forecasting ANNs with no loss, or slight improvement in model performance. Models with fewer inputs have several other favourable characteristics, such as they are less computationally expensive to train, have lower data requirements, and lower complexity (and thus less tendency to become overfitted). Despite these improvements, the input reduced models still tend to underpredict high flows. Additionally, several models have poor PI performance, which may be associated with poor model timing. The topics of model timing and assessing peak flow performance is assessed in the following chapter, which proposes two procedures for improving peak flow and utilizes specialized performance measures to characterize error in terms of amplitude and timing.

CHAPTER 4. ASSESSING THE EFFECTS OF ERROR WEIGHTING AND BOOSTING USING VISUAL PERFORMANCE MEASURES FOR ARTIFICIAL NEURAL NETWORK-BASED FLOOD FORECASTING

4.1 Introduction

Data-driven approaches such as artificial neural networks are increasingly being used for flood forecasting applications (ASCE 2000a, 2000b). Such models are simple to develop and often outperform conventional, physically-based approaches (Mosavi et al. 2018). A common operational use of flood forecasting models is for flood early warning systems (EWS) (Yilmaz et al. 2010, Islam and Islam 2010, Kauffeldt et al. 2016). It is important to consider this application during model development and evaluation (Bennett et al. 2013). In the case of EWS, the model accuracy (both the timing and amplitude) is very important during large hydrological events, and much less important during low flow events. Thus, by necessity EWS models should be calibrated to have a higher efficacy during large events.

4.1.1 Peak error characterization

Several recent studies have made a distinction between timing (sometimes called phase, temporal, or horizontal) and amplitude (sometimes called magnitude, and vertical) error and developed approaches for decomposing error into these two components (Liu et al. 2011, Ewen 2011, Ehret and Zehe 2011, Seibert et al. 2016). The need to characterize error in such a way, compared to typical error measures such as NSE, is illustrated in Figure 4-1 below; this visualization shows how positive and negative timing and amplitude error for predictions that all have the same NSE value, which is sometimes called equifinality (Gupta et al. 2009, Liu et al. 2011). Relying on one composite metric (such as the NSE) that does not distinguish between timing and amplitude may result in over-confidence of model performance in the context of EWS models.

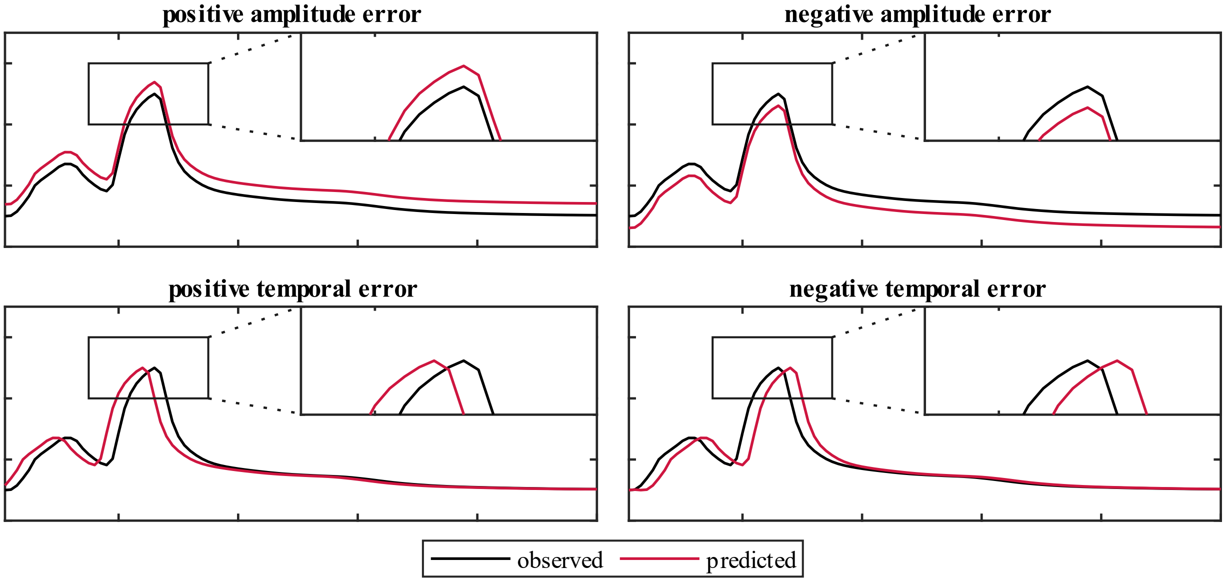

Figure 4-1: Equifinality between positive and negative timing and amplitude errors; each case has an NSE value of 0.8346.

For the purpose of EWS, the sign of each timing and amplitude error is very important. Positive timing error (early prediction) and negative amplitude error (overprediction) are generally less consequential than negative timing error (late prediction) and positive amplitude error (underprediction), since EWS models typically attempt to predict significant flood events prior to their occurrence. For ANN-based EWS, timing sometimes attributable to the calibration procedure and cost function, which is described in the following section.

4.1.2 Model calibration

This section outlines common calibration procedures for flood forecasting ANN models, followed by two proposed modifications to standard calibration practices.

4.1.2.1 Common calibration procedures

ANN models for flood forecasting are typically calibrated using first- or second-order local, deterministic methods; backpropagation (BP) is the most widely used class of calibration algorithms for ANN models (Maier et al. 2010). This class of algorithm has been very successful because of its speed and accuracy. However, many of the most powerful and popular learning algorithms are limited to cost functions that consider types of error that are based on the difference between points with the same abscissa (i.e., amplitude-based error) such as the sum of squared errors (SSE) or mean squared error (MSE) (Seibert et al. 2016), due to their use of the Jacobian matrix. In other words, the backpropagation algorithms are explicitly designed to minimise the amplitude error of predictions rather than the timing error (or both).

This contributes to predictions exhibiting timing error, which is common in ANN-type models (Conway et al. 1998, Abrahart et al. 2010). Such models tend to develop an overreliance on autoregressive inputs, which may be attributable to the amplitude-type cost functions being insensitive to minor timing errors (De Vos and Rientjes 2005, Abrahart et al. 2010, Ehret and Zehe 2011). Previous research has demonstrated that using root mean squared error (RMSE) for model calibration does not consistently produce a model with an optimum timing error, and that timing error may be reduced at the cost of a larger RMSE (De Vos and Rientjes 2005, Abrahart et al. 2010). There are two important issues related to ANN calibration procedures: first, by relying on the standard backpropagation calibration algorithms, the models are restricted to using amplitude-based cost functions, which in turn result in timing errors; second, improved timing performance of ANN models may be achieved at the expense of lower performance in terms of amplitude. Therefore, typical calibration procedures that rely on amplitude-type cost functions are ineffective for minimizing timing errors, hence there is need for modified calibration approaches.

The following section proposes two approaches for improving the timing of ANN models, both of which are constrained to using calibration procedures that are limited to using amplitude-type cost functions; such cost functions are most commonly and widely used in ANN models.

4.1.2.2 Correction procedures

Correction procedures are approaches to adapting ANN calibration to improve the timing performance of predictions. Abrahart *et al.* (2010) apply a correction procedure originally proposed by Conway *et al.* (1998) to improve the timing performance of flood forecasts; the approach uses a neuro-evolutionary technique that penalizes models, such that they are 'bred out' from the population. The approach was able to improve the timing for models with short lead times, however models with longer lead times saw little improvement (Abrahart et al. 2010). Neuro-evolutionary approaches provide more flexibility for different cost functions compared to backpropagation-based calibration, however, do not match the speed or accuracy of backpropagation-based approaches. This research proposes two correction procedures that are based on backpropagation-type calibration: error weighting and boosting.

Error weighting simply involves weighting residuals during the model calibration phase to place greater emphasis on high flows or high gradients, which is typically when high timing error is observed (i.e., during the rapid rising or descending limb of a hydrograph). Four different error weighting schemes are used in this research, which are described in section 4.2.2.1.

Boosting iteratively trains a series of models that are each trained to predict the residuals of the previous model and the boosted prediction is given as the weighted sum of the series of models, which is discussed further in section 4.2.2.2. Boosting is a promising technique for improving peak flow accuracy, as it is well suited for fitting extreme values (Ridgeway 1999).

4.1.3 Model performance assessment

There are a wide variety of performance measures that have been developed for evaluating the performance of hydrological predictions (Bennett et al. 2013). While evaluating model performance, it is widely considered good practice to use several measures, as different measures will capture different model characteristics (Moriasi et al. 2007, Maier et al. 2010, Khan and Valeo 2016b). The various measures have been summarized and categorized in several review papers (Maier et al. 2010, Bennett et al. 2013). The following sections describe the limitations of standard (i.e., commonly used) performance measures and introduce visual measures, a class of performance measures intended to replicate an expert's visual assessment of the agreement between two hydrographs.

4.1.3.1 Standard performance measures

RMSE and NSE, which are used to assess the models in the previous chapter, are among the most widely used for environmental models (Gupta et al. 2009, Maier et al. 2010, Bennett et al. 2013). However, despite their widespread use, overreliance on these measures may result in poor characterization of model performance, especially for assessing peak flow performance.

These common performance measures are frequently criticized in the literature. For example, ANN models for sunspot prediction produced a lower RMSE compared to conventional models, however were found to have no predictive value (Abrahart et al. 2010). Similarly, NSE values may be misleadingly favourable if there is significant observed seasonality (Ehret and Zehe 2011). NSE is also associated with the underestimation of large peak flows, volume balance errors, and undersized variability (Gupta et al. 2009, Ehret and Zehe 2011). Ehret and Zehe (2011) evaluate the relationship between phase error and RMSE using triangular hydrographs; this study shows how RMSE is highly sensitive to minor phase errors, however, when a hydrograph has a phase and amplitude error RMSE is much more sensitive to overpredictions compared to underpredictions.

Despite being frequently criticized, the popularity of measures such as NSE and RMSE persists in studies and practical applications of flood forecasting models (Seibert et al. 2016). Typically, little justification is included for the use of performance measures, beyond stating their widespread use in hydrological analysis.

4.1.3.2 Visual performance measures

Visual performance measures (VMs) are a class of measures intended to quantify the judgement of a hydrologist comparing the differences between two hydrographs (Bennett et al. 2013). While such measures cannot replace expert judgement, they can be used to characterize error as intuitive, physically-based dimensions (timing and amplitude). This is useful for diagnosing performance issues and understanding physical behaviour within watersheds (Crochemore et al. 2014).

This chapter evaluates three different VMs: peak difference (PD), series distance (SD), and hydrograph matching (HM), which are illustrated in Figure 4-2. PD is the simplest of the three measures and is given by the Euclidian distance between observed and predicted peak flows for each hydrological event. SD involves segmenting observed and modelled events into rising and falling peaks, then comparing the Euclidian distance between pairs of polyline segments (Ehret and Zehe 2011, Seibert et al. 2016). HM involves drawing rays between observed and modelled points, while permitting for a set number of repetitions and skips between points, and minimizing the sum of the rays based on a time versus amplitude weighting parameter (Ewen 2011).

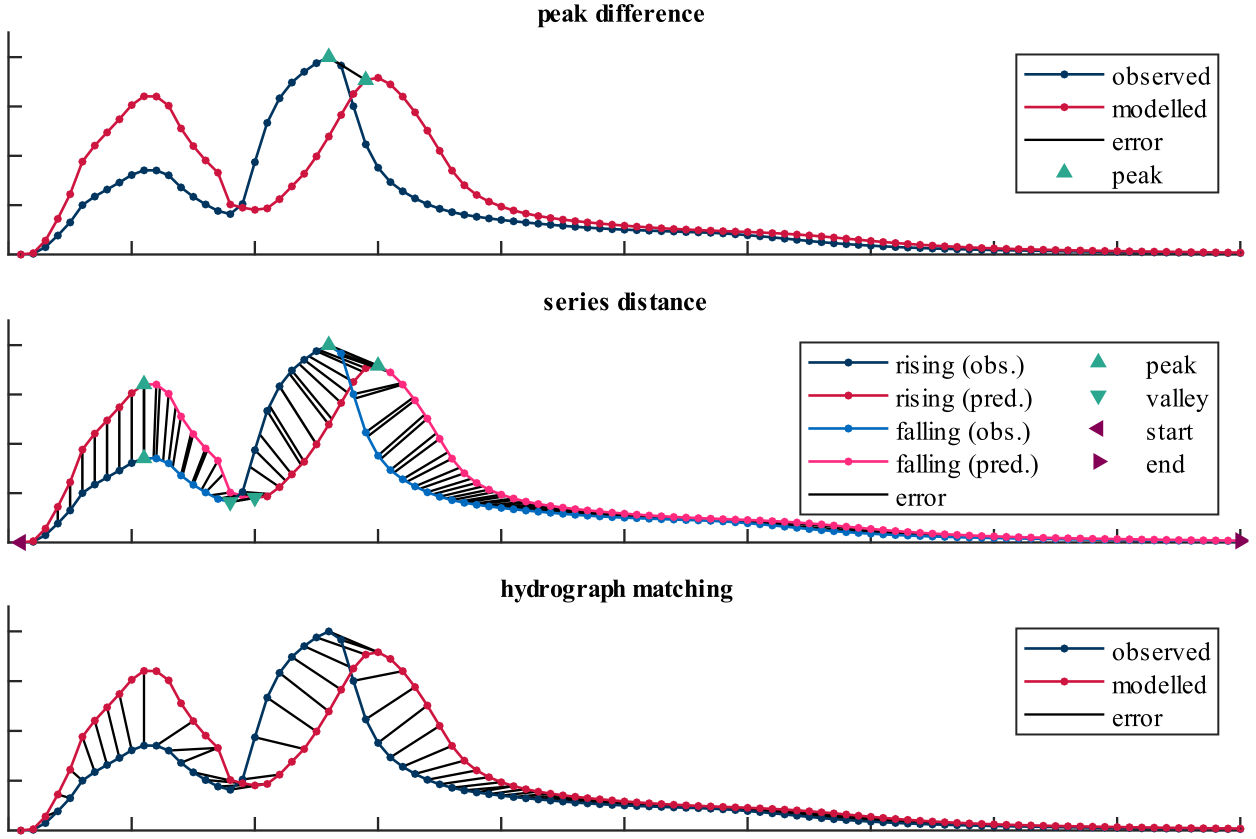

Figure 4-2: Visual example of the three visual performance measures used in this study.

All three methods have multiple tuning parameters that are manually adjusted such that they produce amplitude and timing estimates that agree with a visual assessment of the observed and modelled hydrographs. Also, while the VMs could be used on a continuous simulations, in this research they are used on an event basis as to evaluate each event independently, because different events may have different timing errors (Seibert et al. 2016).

4.1.4 Objectives

The objective of this research is to evaluate different approaches of adapting the ANN calibration procedure to correct the timing error for large hydrological events. The effectiveness of each correction procedure is evaluated using standard performance measures, on a continuous and event basis, and VMs. The objectives of the performance assessment are to study the effects of the various correction procedures and to determine whether VMs are able to quantify model characteristics not captured by standard measures.

4.2 Methods

The following section describes the correction procedures, performance assessment, and performance visualization used in this study.

4.2.1 Model configuration

Table 4-1 below summarizes the hyperparameters used for the baseline ANN model. The parameters are quite typical of models used for flood forecasting (ASCE 2000a, Maier et al. 2010, Khan and Valeo 2016b, Khan et al. 2018). The hidden layer size was determined based on a simple grid search, based on MSE. The difference in performance between 10 and 40 neurons is minimal, however the hidden layer was sized sufficiently large such that it does not limit the ability for the model to utilize exogenous inputs. The Levenberg-Marquardt backpropagation (LMBP) calibration algorithm is used in this research, due to its popularity and it was found to outperform the other calibration algorithms based on speed and MSE (Maier et al. 2010).

Table 4-1: ANN hyperparameters and experimental setup

	Parameter	Value
Architecture	ANN Type	Multi-layer perceptron
	Input nodes	120
	Hidden nodes	20 (single layer)
	Output nodes	1
	Activation functions	Hyperbolic tangent(hidden), Linear (output)
	Normalization	
Data partitioning	Partition style	Block (2-years)
	Calibration[1]	80% (2000-2008)
	Testing	20% (2009-2010)
	Validation method	KFCV (four 2-year folds)
Calibration	Algorithm	LMBP
	Cost function	MSE
	Early-stopping	Validation stopping (6 epochs)
	Iterations[2]	20-100

[1] Includes training and validation data

[2] Fewer iterations for boosting models due to higher computation time

4.2.2 Correction procedures

The following sections describe the methods used for the error weighting and boosting correction procedures. The first, error weighting, is an obvious approach for improving peak performance, yet is not described in literature for ANN-based flood forecasting applications. The second procedure, boosting, is commonly used in classification models. While some studies evaluate the effects of boosting on flood forecasting models, there is no research explicitly studying its effects on peak flowrate performance.

4.2.2.1 Error weighting

Error weighting (EW) involves weighting the residuals of individual predictions for calibration or assessing performance. In the context of flood forecasting, weighted error may be used to place greater emphasis on high flows during training, as to improve accuracy during large hydrological events at the expense of performance during average or low flow conditions. Conversely, this method could be also adopted for applications such as drought prediction where the low flow conditions are weighted higher than the high flows.

For calibration, EW is commonly used for weighted least squares (WLS) models, which are linear regression models that typically use error weights to counter the influence of heteroscedasticity during model calibration by weighting samples by the reciprocal of variance (Almeida et al. 2002, Strutz 2015). EW is easily adaptable for to the calibration of ANN-type regression models.

To assess model performance, Pauline, See, & Smith (2001) use RMSE weighted by observed flow and gradient (change in flow from previous timestep). The weighting schemes each have their limitations; weighting based on observed flow value under-prioritizes the beginning and ending of events, whereas weighting based on flow gradient under-prioritizes sustained high flows. The weighted error schemes are proposed as performance measures that are relevant to EWS and used in a comparison between different ANN architectures, not for weighting the ANN cost functions (Pauline et al. 2001). Bennet et al. (2013) provide a summary of performance measures for environmental modelling; information weighting is included amongst these, which has been demonstrated as superior to uniformly weighted error for comparing images versus distorted variations (Tompa et al. 2002). These various examples of weighted error for assessing performance provide the precedence for the cost function weighting schemes proposed in this research.

In this chapter, four different error weighting methods are considered, based on: by applying a linear transformation to observed flow, a logistic transformation, the flow gradient, and the information value of the flow. Linear weighting simply weights the error based on the magnitude of the observed flow value, such that errors occurring at high flows receive greater weight in the cost function. Logistic weighting is tuned such that high flows receive a very high weighting, whereas low flows receive almost no weight. Gradient weights errors based on the absolute flow gradient, which adds weight to the rising and falling limbs. Finally, information weighting places more emphasis on flows with high entropy, which is approximated as improbable flows. Distinct from the other weighting schemes, error weighting prioritizes infrequent flows, whether high or low, and less importance is placed on frequent flows. These four schemes are summarized in Table 4-2 below, and visualized in Figure 4-3: Normalized observed flow (black line) and four error weight schemes plotted temporally (top row) and sorted by flow magnitude (bottom row)

Table 4-2: Equations for four different error weighting schemes used in this research.

Name	Equation	Equation number
ew_linear	$\mathrm{ew}_{\mathrm{t}} = \mathrm{norm}(\mathrm{q}_{\mathrm{t}})$	(18)
ew_logistic	$\mathrm{ew}_{\mathrm{t}} = \mathrm{norm}\left(\frac{1}{1 + \exp(\alpha(\mathrm{q}_{\mathrm{t}} - \bar{\mathrm{q}}))}\right)$	(19)
ew_gradient	$\mathrm{ew}_{\mathrm{t}} = \mathrm{norm}(\lvert \mathrm{q}_{\mathrm{t}} - \mathrm{q}_{\mathrm{t}-1} \rvert)$ $\mathrm{ew}_{\mathrm{t}} < \beta = \beta$	(20)
ew_information	$\mathrm{ew}_{\mathrm{t}} = \mathrm{norm}\left(\log\left(\frac{1}{\mathrm{P}(\mathrm{q}_{\mathrm{t}})}\right)\right)$	(21)

The term ew_t is an array of weights that is the same size as the observed flow, q_t. The parameter α for the logistic weighting in equation (19) is used to tune the inflection point that separates low (less important) flows from high flows (more important); a value of -16 was found to produce a reasonable weighting scheme. The parameter β in equation (20) is an optional parameter used to ensure that error weights are not equal to 0; the gradient-based error weights tend to have a high proportion of values equal or near to 0, which may result in poor model performance during steady flows. This research uses a β value of 0, which is most likely to produce poor predictions during steady flows, due to the low error weight attributed to these flows. Finally, the flow probability

$P(q_t)$ in equation (21) is determined using the automatic histogram binning algorithm in MATLAB 2019a.

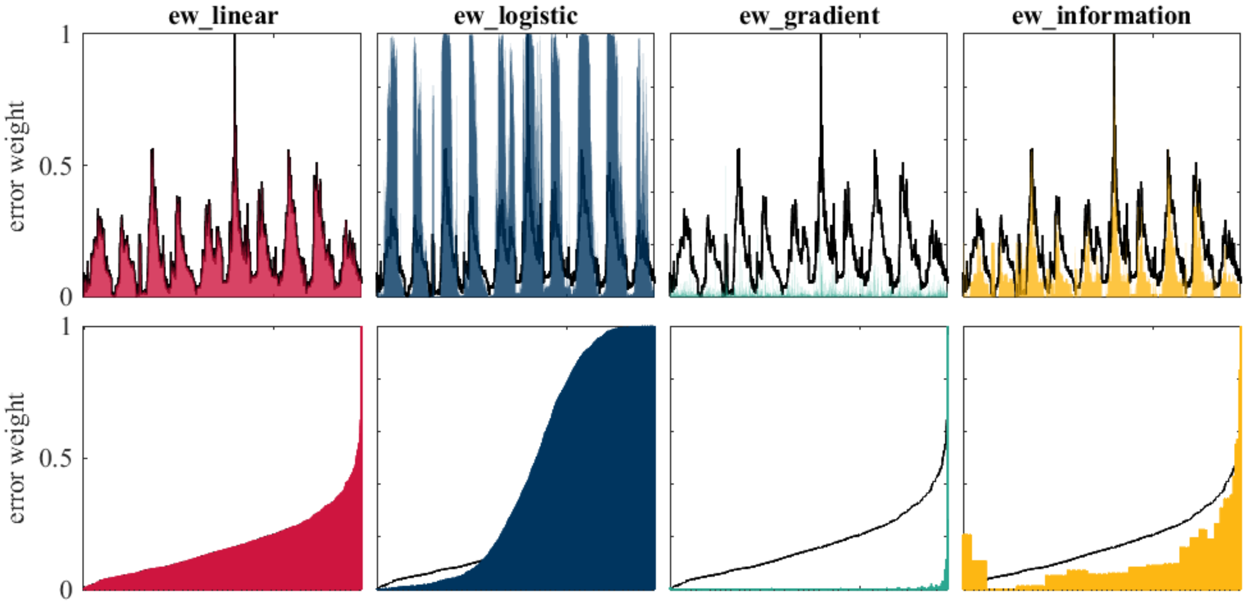

Figure 4-3: Normalized observed flow (black line) and four error weight schemes plotted temporally (top row) and sorted by flow magnitude (bottom row) for linear, logistic, gradient, and information (from left to right) weighting schemes.

The various weighting schemes are visualized in Figure 4-3 that show the normalized observed water level plotted chronologically and sorted by magnitude, along with the corresponding error weight values. In particular, this figure illustrates how ew_logistic applies a low weight to low flows and a large weight to high flows, how ew_gradient is dominated by low gradients thus low weights, and how ew_information places more weight on infrequent flows.

4.2.2.2 Least squares boosting

Boosting is a technique used in machine learning where a strong predictor is replaced by a collection of weak predictors (Schapire 1990). There are a wide variety of types of boosting algorithms, many of which have been refined for specific applications (Ridgeway 1999). Boosting solutions have been used in several studies related to water quality and flood forecasting (Anctil and Lauzon 2004, Belayneh et al. 2016, Li et al. 2016, Barzegar et al. 2018). This study proposes using Gradient Boosting (GB), as it is applicable to continuous output data (whereas some boosting methods only apply to classifiers, where the output is binary) and is well suited for fitting outliers (Friedman 2002).

GB involves training an initial predictor, followed by M subsequent predictors, where each subsequent predictor is trained to predict the residuals of the previous model. The final prediction is calculated as the weighted sum of the collection of (M+1) predictors. This research implements a specific variant of gradient boosting known as least-squares boosting (LSB), where a least-squares cost function is solved at each boosting iteration (J.~H.~Friedman 2000).

The pseudocode for the LSB algorithm is included below. An initial prediction ($\hat{y}_0$) is made using a trained ANN, which is a function of the input set (x) and the observed flow (y). The residuals, e_m (line 3), are predicted, $\hat{e}_m$ (line 4), using a new ANN, using the same input set, x. An adaptive weight, ρ, is calculated based on the minimization of the SSE (line 5), which regulates the boosting process, ensuring that each boosting iteration lowers the overall squared error. A tuning hyperparameter, called the learning rate ($0< \nu \leq 1$), is used to govern the step size of each boosting iteration. Finally, the new prediction, $\hat{y}_m$, is calculated by adding the boosted model, $\hat{e}_m$, is weighted by ρ and ν, to the previous iteration's prediction, $\hat{y}_{m-1}$ (line 6). The process is repeated for M boosts.

The two boosting hyperparameters: the learning rate and the number of boosts, are selected somewhat arbitrarily; however, they are on the low and high end of typical values used in research. Studies often neglect to specify the learning rate, or use a learning rate of 1; while other studies recommend using a low learning rate to reduce the risk of overfitting (van Heijst et al. 2008, Erdal and Karakurt 2013). Therefore, this research considers a low and a high learning rate with values of 0.1 and 1, respectively. Similar to the learning rate, the number of boosts is often unspecified, or a termination criterion is used to determine the number of boosts (van Heijst et al. 2008). This research evaluates the effects of a single boosting iteration, and 5 boosts; the latter was selected as the computational expense of the boosting procedure was on the order of several days and a review of model behaviour suggested that there are minimal benefits of boosting beyond this number.

```
$\hat{y}_0$ = train(x,y)
**for** m = 1 to M
    $e_m = y - \hat{y}_{m-1}$
    $\hat{e}_m$ = train(x,$e_m$)
    $\rho_m$ = argmin($\sum(e_m - \rho_m \hat{e}_m)^2$)
    $\hat{y}_m = \hat{y}_{m-1} + \nu \rho_m \hat{e}_m$
end
```

4.2.3 Performance assessment

The following section outlines the standard measures and VMs, which are used to measure the effects of the various correction procedures.

4.2.3.1 Standard performance measures

In order to assess whether the VMs capture model characteristics that are unrepresented by standard performance measures, we calculate several standard performance measures including NSE and RMSE, provided in the following equations:

$$\mathrm{NSE} = 1 - \frac{\sum(\mathrm{q_t} - \hat{\mathrm{q}}_\mathrm{t})^2}{\sum(\mathrm{q_t} - \bar{\mathrm{q}})^2} \tag{22}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum(\mathrm{q_t} - \hat{\mathrm{q}}_\mathrm{t})^2}{\mathrm{n}}} \tag{23}$$

in which $\mathrm{q_t}$, $\mathrm{\hat{q}_t}$, $\mathrm{\bar{q}}$ correspond to the observed, predicted, and mean observed flows, respectively, and *n* is the number of samples. Note that while NSE and RMSE are both included due to their popularity for hydrological models, their values or mention may be interpreted synonymously as they are both simply the sum of squared residuals between the observed and predicted values, normalized in different ways.

Another standard measure, called the NSE timing (NSET) is also calculated. This measure estimates timing error, but is distinct from the VMs such that it does not attempt to replicate visual comparison of two hydrographs, or calculate localized timing error (De Vos and Rientjes 2005). NSET is conceptually similar to maximum cross-correlation, however, NSET is calculated at each timestep instead of correlation; it has the following equation:

$$\mathrm{NSET} = \underset{\mathrm{L_{min}} \leq \mathrm{L} \leq \mathrm{L_{max}}}{\operatorname{argmax}} \left(1 - \frac{\Sigma(\mathrm{q_t} - \mathrm{\hat{q}_{t+L}})^2}{\Sigma(\mathrm{q_t} - \mathrm{\bar{q}})^2}\right) \tag{24}$$

where *L* is a time shift applied to the predicted flow $\mathrm{L_{min}}$ and $\mathrm{L_{max}}$ correspond to the lower and upper possible timing errors, taken as -4 and 4 in this research, as it is not expected for the timing error to exceed the lead time.

In addition to these measures, two intuitive measures are used to assess if the observed data are captured within the uncertainty envelope of the predictive ensemble ANN models: the coverage and precision. Recall, the uncertainty in this model is owed to the random initialization of ANN parameters and KFCV sampling procedure.

Coverage (CVG), also called Percent Captured, is the fraction of observed samples that fall within the uncertainty envelope ($\mathrm{n_{obs}}$) over the total number of samples (Seibert et al. 2016, Khan and Valeo 2016a). Coverage is between 0 and 1, with 0 corresponding to no samples contained within the uncertainty envelope and 1 corresponding to all the samples falling within the envelope, given as follows:

$$\mathrm{CVG} = \frac{\mathrm{n_{UE}}}{\mathrm{n}} \tag{25}$$

Precision (PRC), also called Prediction Interval Width, is the average difference between the maximum (UE^{+}) and minimum (UE^{-}) predictions (Seibert et al. 2016, Khan and Valeo 2017). The variant of PRC used in this research is normalized by dividing the interval width by the observed flow ($\mathrm{q_t}$) for each timestep, as recommended by Seibert et al. (2016). Lower precision corresponds to a smaller uncertainty envelope, with a value of 0 corresponding to no uncertainty, given as follows:

$$\mathrm{PRC}^{*} = \frac{1}{\mathrm{n}} \sum \frac{\left(\mathrm{UE}_{\mathrm{t}}^{+} - \mathrm{UE_t}^{-}\right)}{\mathrm{q_t}} \tag{26}$$

Both coverage and precision are important measures for characterizing performance. Coverage is especially important in cases where the model uncertainty is used to generate probabilistic predictions. For example outputting model predictions as the percentage of exceeding a certain flow level, rather than a discrete value such as the ensemble mean.

4.2.3.2 Visual performance measures

Each of the three specialized performance measures described in the following sections has been adapted to quantify error as timing and amplitude components. Authors of these methods have proposed alternative metrics based on the same methodology (e.g., skill); however, these metrics are not evaluated in this study as they are less easily compared with each other, or with a visual inspection of the respective observed and modelled hydrographs (Ehret and Zehe 2011).

4.2.3.2.1 Automated hydrological event identification

Hydrological events are identified by first thresholding flows based on a critical percentile value, then selecting prominent events using the *findpeaks* MATLAB function. This function has been used for peak identification in several studies (c.f. Manfreda et al. 2018, Blaszczak et al. 2019). This research uses a threshold value at the 90th percentile flows and selects events with a minimum peak prominence of 0.3 m. Five events meet these criteria throughout the 10 years of data. The start and end times for events corresponds to the point at which it crosses the static threshold value; the peak corresponds to the maximum value and corresponding timestep within each event.

The method described above is relatively simple, more sophisticated methods have been proposed that use data-driven methods for identifying hydrological events, which may be implemented in future research (Thiesen et al. 2019).

While VMs may be adapted for use on continuous data, different predicted hydrological events may have different timing errors, therefore, VMs are calculated on an event basis. Standard performance measures are also calculated on an event basis, for the sake of comparison.

The following three sections outline the procedures for calculating the three VMs: PD, SD, and HM.

4.2.3.2.2 Peak Difference

The Peak Difference (PD) metric compares discrete modelled observed and peak flows. Observed peak flows are identified as the highest magnitude flows within each hydrological event (line 2) and modelled peaks are calculated as the largest predicted values within a specified proximity to the observed peak (line 3). The search proximity is a hyperparameter and denoted by w1 and w2, corresponding to the number of timesteps to the left and right respectively. In this research, the search proximity is set equal to the lead time (4 timesteps) in both directions, as it is very unlikely that the timing error between the observed and predicted peaks exceed this window. The PD values for timing and amplitude are calculated as the difference between the peak flows (line 4) and the difference in time indices at which the peaks occur (line 5).

```
1   **for** event = 1 to num_events
2   ....$[P_{obs},i] = \max(obs_i)$
3   ....$[P_{mdl},j] = \max(mdl_{i-w1} \ldots mdl_{i+w2})$
4   ....$PD_a = P_{obs} - P_{mdl}$
5   ....$PD_t = i - j$
6   **end**
```

4.2.3.2.3 Series Distance

The following method describes an adaptation of the SD method first described and later improved in a series of papers (Ehret and Zehe 2011, Seibert et al. 2016). This method firstly identifies critical hydrological points in the observed hydrograph including peaks and valleys (lines 2-3). Matching predicted critical points are calculated using a simple search window (line 4). Both the observed and predicted hydrographs are partitioned based on the critical points and the partitions are classified as either rising limbs or falling limbs (line 6). The Euclidian between each segment pair (observed and predicted) are calculated and decomposed into timing and amplitude components (lines 7-9). Finally, the SD value for the event is given as the mean between all the segments (lines 10-11). The SD could be calculated as the weighted mean (e.g., weighting rising limbs higher than falling limbs) if desired; segments are not weighted in this research.

The method described for pairing observed and predicted hydrograph segments in this research is different than that described by Seibert *et al.* (2016), which uses a coarse-graining approach to iteratively dissolve segments until observations and predictions are attenuated. The method used in this research is less robust, but is found to generate reasonable segment pairs, has lower computational requirements, and does not discard any data.

Another distinction from the original SD method is the peak identification method. The original method recommends smoothing noisy data prior to identifying peaks, while the implementation in this research uses the *findpeaks* function in MATLAB. The *findpeaks* function eliminates the need for data smoothing; smoothing can be problematic as it does not always preserve the timing of peaks (the timing of the peak is an essential component for measuring ANN performance in the context of EWS models).

One improvement made in this research is that the calculation of the Euclidian distance between segments calculates the number of vertices based on the least common multiplier between the number of observations and predictions in each segment, rather than only the number of points in the observed segment, thus ensuring no predicted values are omitted from the distance calculation. Each observed peak and valley is paired with a with the maximum and minimum predicted point, respectively, that falls within the search window of -4 to 4 timesteps (see section 4.2.3.2.2).

```
for event = 1 to num_events
........$p_{obs}$ = find_minor_peaks(obs,pro)
........$v_{obs}$ = find_minor_valleys(obs)
........$p_{mdl}$ = match_peaks($p_{obs}$,w1,w2)
........$v_{mdl}$ = match_valleys($v_{obs}$,w1,w2)
........classify_hydrograph_segments(p,v)
....for sgmt = 1 to num_segments
........[x,y] = calc_polydist($sgmt_{obs}$,$sgmt_{mdl}$)
....end
....$SD_a$ = mean(Y)
....$SD_t$ = mean(X)
end
```

4.2.3.2.4 Hydrograph Matching

The HM algorithm is a visual measure proposed by Ewen (2011). Similarly to the SD method, HM involves drawing rays between observed and predicted hydrographs, however, unlike SD, this method does not partition the events into segments (i.e., rising and falling limbs); instead, all possible ray locations are considered and the set of rays corresponding to the minimum cumulative ray length is selected (Ewen 2011).

```
def erfun(obs,mdl,j,k) = ([obs(j) – mdl(k)]$^2$ + b$^2$[j – k]$^2$)

for 1 to num_events
....J = num(mdl)
....K = num(obs)
....$C_{1:J,1:K,1:2}$ = 9999
........for j = 1 to 1+w1
............k = 1
............e = erfun(obs,mdl,j,1)
............Cj,1,1 = e;
........end
............for k = 2 to K
................for j = k-w2 to k+w1
....................if j<1 or j>K
........................next cycle
....................end
....................e = erfun(obs,mdl,j,k)
....................$C_{j,k,2}$ = e + $C_{j,k-1,1}$
....................if j>1
........................m = min($C_{j-1,k-1,1:2}$)
....................end
....................if j>2
........................m = min(m, $C_{j-2,k-1,:}$)
....................end
....................$C_{j,k,1}$ = e+m
................end
............end
ray_indices = min(min($C_{:,:,:}$,3),1))
$HM_a$ = obs – mdl(ray_indices)
$HM_t$ = 1:K – ray_indices
end
```

The algorithm systematically considers ray positions while following strict rules (e.g., two rays may be connected to a single point). The optimum ray locations are selected based on the minimum cumulative distance across the entire hydrological event. This approach was developed based on the more general Minimal Variance Matching (MVM) algorithm (Latecki et al. 2005). The pseudocode included above is simple but not intuitive to interpret and therefore not described line-by-line as was provided for the other methods; a more detailed description is provided in the original paper by Ewen (2011). The implementation in this research has no functional changes aside from the calculation of timing and amplitude error appended (lines 28-30).

4.3 Results and discussion

The following sections present and discuss the model results for the chapter. Firstly, the baseline model results are presented, including their VM performance. This is compared to the corrected models across standard and VM performance measures, followed by discussions on the efficacy of each correction procedure and the VMs.

4.3.1 Baseline standard performance

The baseline model, which is calibrated using the unmodified LMBP algorithm (i.e., without any correction procedures) is used for comparison between the various correction procedures. The baseline model has a very strong median NSE of 0.969 but exhibits poor timing performance. The deficiency of the baseline model is illustrated in the correlation plots and hydrographs in Figure 4-4 and Figure 4-5, respectively. The cluster of points in the top right of the subplots in Figure 4-4, show that the predicted water levels are below the 1:1 line of perfect fit, indicating that the model underpredicts high flows. Underprediction of high flows is further demonstrated by the predicted hydrograph shown in Figure 4-5, which exhibits clear timing error, with the rising limb of the June 18 2005 event, which is not contained within the 99% confidence bounds.

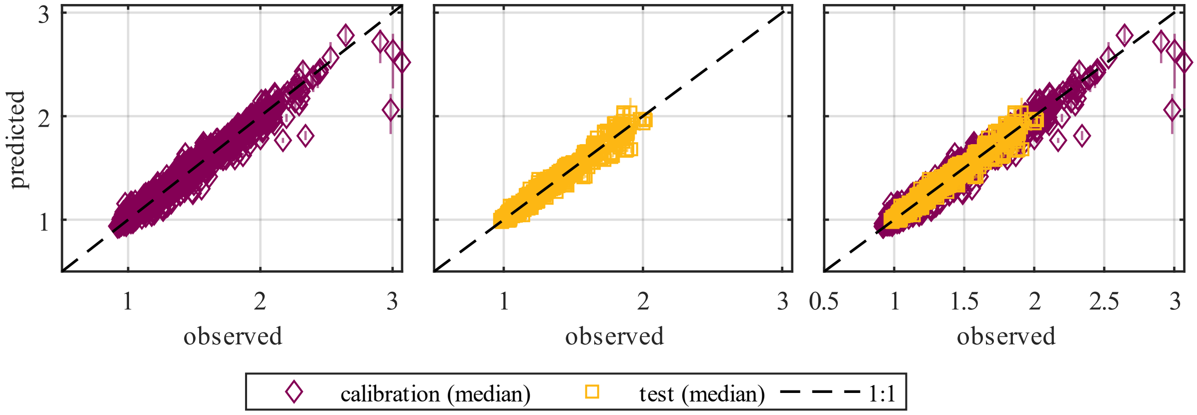

Figure 4-4: Left: calibration (purple) predictions versus observed flow levels. Middle: test (yellow). Right: calibration and test data. Vertical bars indicate the quartile range of the predicted ensemble.

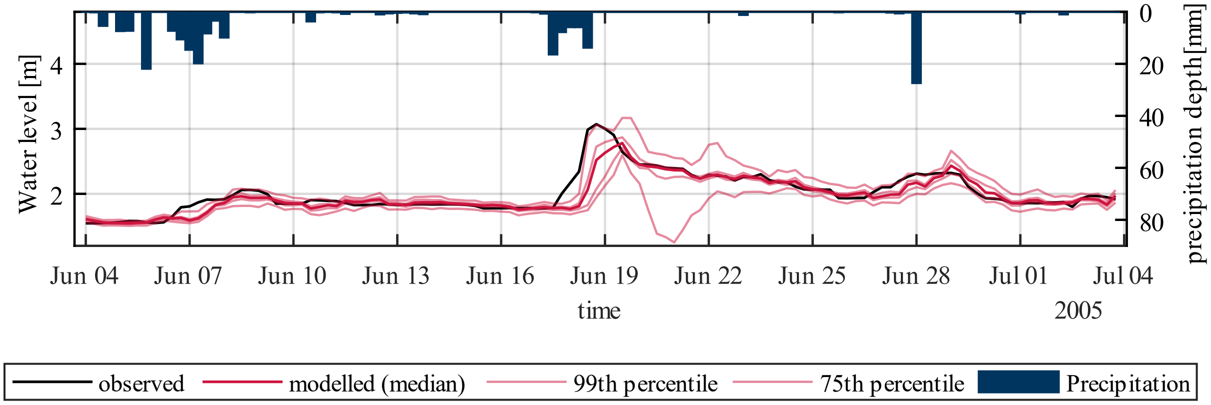

Figure 4-5: Observed and predicted water level, including prediction 75% and 99% confidence envelopes.

4.3.2 Baseline visual performance

The visualizations in Figure 4-6 below (and Figure 4-8 in the following section) are called ‘peak-boxes’ and used to illustrate timing and amplitude error of the various VMs (Zappa et al. 2013).

While error values of 0 are ideal, positive timing and negative amplitude errors are typically less consequential compared to negative timing and positive amplitude for EWS models.

The five largest events from the 10-year period are shown in Figure 4-6, for a sample model from the baseline ensemble, along with corresponding visual performance values. In the spirit of VMs, the accuracy of the VMs can be judged by comparing the amplitude and timing error estimates to the observed and predicted hydrographs in the top row of Figure 4-6 (Bennett et al. 2013, Seibert et al. 2016). The error estimates made by the VMs are typically in agreement with each other and discussed in greater detail below. The VMs tend to underestimate the timing error, perhaps because the eye is drawn to timing difference between the observed and predicted rising limb, where the difference in timing is clearest, compared to during steady flows.

Based on the hydrograph, the prediction for event 1 appears to have minimal timing and amplitude error, which agrees with SD and VM. The PD estimates a timing error of 1 timestep simply because the predicted rising limb continues to rise while the observed rising limb peaks and plateaus. Event 2 has a considerable timing error and severely underestimates the peak amplitude, which is best captured by the PD measure. SD and HM estimate lower timing and amplitude error because the peak error becomes averaged out throughout the remainder of the event, which has lower visible timing and amplitude errors.

Event 3 exhibits similarly poor timing, however unlike for event 2, the timing error is not captured by the PD measure because the observed peak occurs after a high flow plateau lasting several timesteps, and subsequently being aligned with the delayed predicted rising limb. Hence SD and HM are slightly more representative of the timing error for this event; however, SD is biased by the same deficiency, such that the partitioning of the rising and falling limbs occurs at the right of the observed plateau and the left of the predicted plateau. Event 4 has 3 distinct peaks, with a clear timing error on two peaks and relatively low amplitude error. This error is well represented by the three VMs, which collectively estimate a negative timing error between 0.5-1.5 timesteps, and little amplitude error. Lastly, event 5 has a more pronounced timing error, which is also well quantified by the VMs, which estimate a negative timing error between 1-2.5 timesteps, and minimal amplitude error.

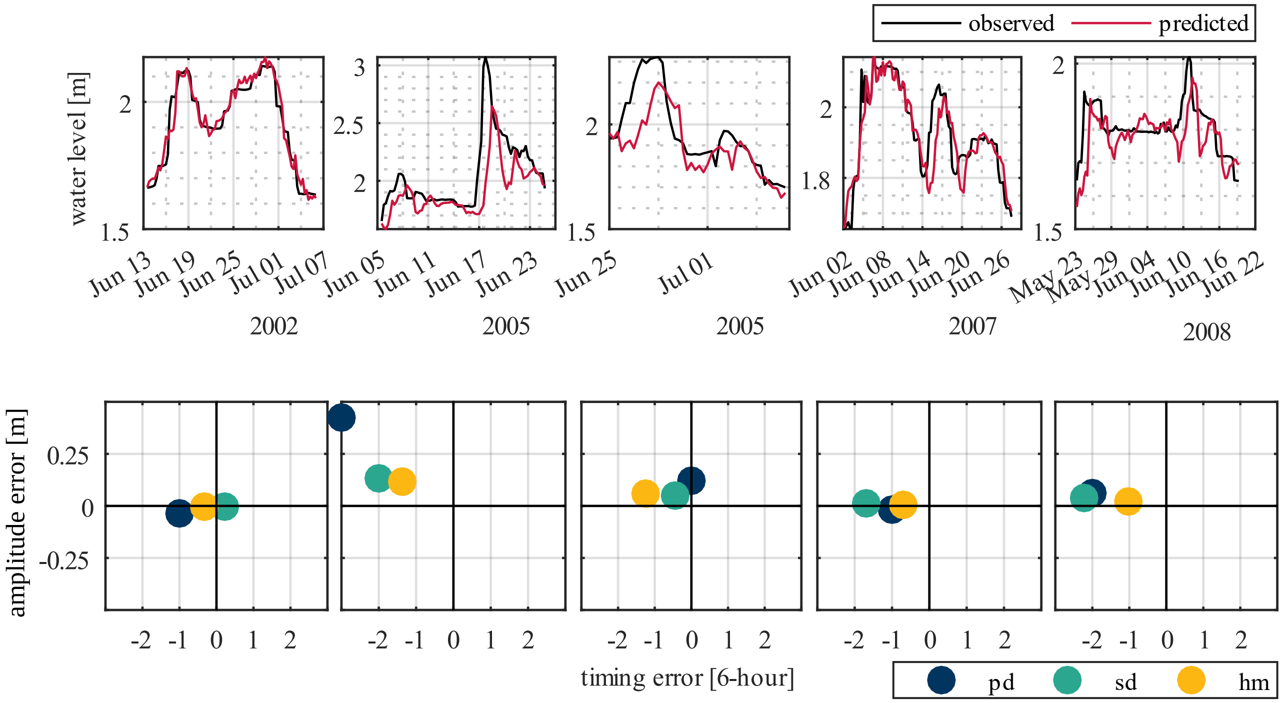

Figure 4-6: Top row: Largest 5 observed events (black lines) and baseline sample prediction (red lines). Bottom row: PD (blue), SD (green), and HM (yellow) amplitude and timing error for each of the 5 events.

4.3.3 Corrected model standard performance

The performance of the baseline model and 8 correction procedure configurations are shown in Table 4-3 below for the calibration and test datasets. The strongest performance scores between configurations for each standard measure are bolded. The baseline ensemble has a very strong

NSE, however, it does not provide any indication of timing error. NSET indicates very poor timing, of -2.6 and -4.0 for calibration and testing, respectively.

Table 4-3: NSE (mean), RMSE (mean), TE (mean) CVG, and PRC for base, errorweighted, and boosted models for the calibration (left columns) and test (right columns) datasets.

	NSE		**RMSE**		**NSET**		**CVG**		**PRC**	
	cal.	test	cal.	test	cal.	test	cal.	test	cal.	test
baseline	0.969	0.968	0.046	0.042	-2.6	-4.0	0.918	0.896	0.111	0.109
ew_linear	0.959	0.950	0.052	0.052	-2.2	-4.0	0.949	0.929	0.152	0.153
ew_logistic	0.955	0.952	0.055	0.052	-2.4	-4.0	0.902	0.875	0.137	0.140
ew_gradient	0.946	0.899	0.060	0.074	**-1.1**	-4.0	**0.973**	**0.967**	0.173	0.174
ew_information	0.946	0.933	0.057	0.059	-2.0	-4.0	0.977	0.966	0.219	0.221
lsb_0.1-1*	0.969	**0.971**	0.046	0.041	-2.8	-4.0	0.687	0.708	0.054	0.054
lsb_0.1-5	0.974	0.970	0.042	0.041	-2.1	-4.0	0.621	0.618	0.046	0.038
lsb_1-1	0.974	**0.971**	0.043	**0.040**	-2.3	-4.0	0.574	0.588	**0.036**	**0.035**
lsb_1-5	**0.977**	0.970	**0.039**	0.041	-1.8	-4.0	0.562	0.559	0.047	**0.035**

*The **lsb** identifiers include the learning rate used, followed by the number of boosts (learning rate – number of boosts)

The EW models typically exhibit a slightly lower NSE compared to the baseline. This is expected; weighting the squared residuals will decrease performance measures based on squared residuals such as NSE and RMSE (which uniformly weight samples). The EW models also tend to have improved NSET performance, but this improvement is only observed for the calibration dataset. Finally, CVG improves significantly for the EW models, whereas PRC decreases, indicating that the EW models have a larger prediction envelope that better contains the observed flow values.

The LSB models have improved NSE and RMSE, which is consistent with recent research on boosting for hydrological models (Belayneh et al. 2016, Barzegar et al. 2018).The NSET is also improved, however, similarly to the baseline and EW models, this improvement is not observed for the test dataset. Finally, the CVG for the LSB models is significantly poorer compared to the baseline, however, the precision is much higher.

Of the various EW and LSB configurations, the ew_gradient and lsb_1-1 configurations are chosen for a more detailed analysis in the following sections, as these were found to have the most distinct performance when compared to the baseline model, and generally performed better according to the VMs. Additional results showing the VM performance for all of the EW and LSB configurations are included in the Appendix.

Table 4-4: Event-based standard mean performance measures, including NSE, RMSE, NSET, CVG, and PRC, for baseline, ew_gradient, and lsb_1-1 model ensembles.

	event	**NSE**	**RMSE**	**NSET**	**CVG**	**PRC**
baseline	1	0.877	0.056	-0.550	0.945	0.115
	2	0.582	0.190	-2.130	0.859	0.180
	3	0.722	0.094	-1.560	0.909	0.131
	4	**0.539**	**0.080**	-1.300	0.922	0.140
	5	-0.612	0.089	-3.720	0.789	0.115
ew_gradient	1	0.653	0.091	-0.220	**0.978**	0.220
	2	0.577	0.184	**-0.950**	**0.936**	0.293
	3	0.409	0.133	**-0.690**	**1.000**	0.238
	4	-0.445	0.138	**-0.270**	**0.981**	0.292
	5	-8.266	0.203	**-1.310**	**0.927**	0.277
lsb_1-1	1	**0.897**	**0.051**	**-0.050**	0.835	**0.066**
	2	**0.631**	**0.176**	-1.950	0.718	**0.110**
	3	**0.766**	**0.085**	-1.100	0.659	**0.089**
	4	0.504	0.079	-0.800	0.738	**0.086**
	5	**-0.601**	**0.089**	-3.850	0.661	**0.074**

Since VMs are calculated on an event basis, standard performance measures are also calculated for each event, which are shown in Table 4-4. The best performance measures are highlighted in bold on an event basis (e.g., the highest NSE for event 1 between the 3 different configurations is bolded). Comparing these methods of assessing model performance will help determine whether assessing performance using VMs adds value, or simply assessing performance event by event is sufficient for identifying model deficiencies.

Evaluating model performance using standard measures on an event basis reveals that the model performs poorly based on standard measures such as NSE, which is consistently lower comparative to the values calculated for the entire dataset. This is because it eliminates the bias introduced by seasonality, which has a period much shorter than the typical length of a hydrological event. Also, the worse performance may be attributable to the tendency for the ANN to have poorer accuracy during high flows, compared to typical, steady flows. Other standard measures typically agree with the values calculated on the entire dataset, such as strong CVG and poor PRC for the EW model, and poor CVG yet strong PRC for the LSB model.

4.3.4 Corrected model visual performance

The timeseries predictions for the baseline, error weighted, and boosted models are shown on an event basis in Figure 4-7 and the VM performance in Figure 4-8.

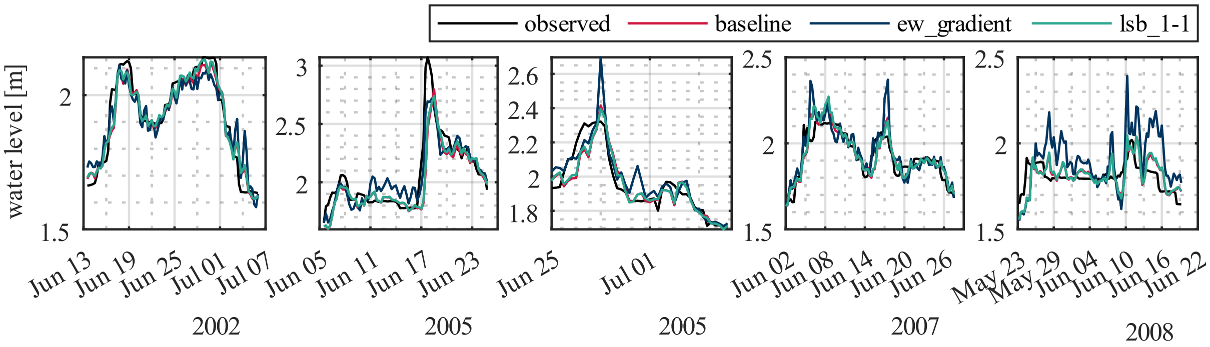

Figure 4-7: Observed hydrological events and mean predicted baseline (red), information weighted (blue), and boosted (green) models.

The ew_gradient model tends to have better timing compared to the baseline according to the PD, SD, and HM (although less dramatic for SD and HM). For 2 out of the 5 events the ew_gradient model has a positive PD timing error, indicating that the models predicted the peak occurring before the observed peak. This may be attributable to the high error weighting Also, as discussed earlier, a positive timing error is typically less consequential than a negative timing error. Additionally, PD indicates a drastic overprediction in terms of amplitude error, which, similarly to a positive timing error, is less consequential than a positive amplitude error.

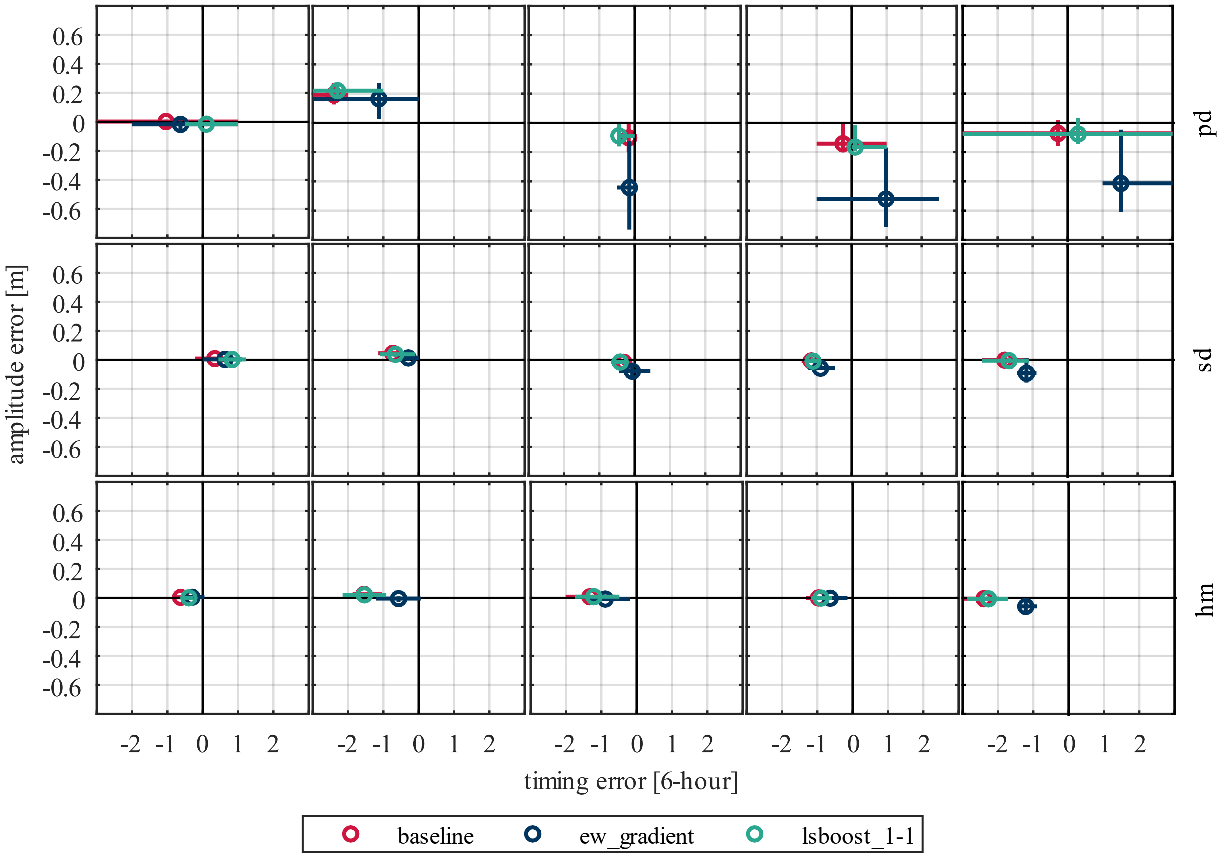

Figure 4-8: Performance of visual measures, PD (top row), SD (middle row), HM (bottom row), for baseline (red), information weighted (blue), and boosted (green) models. Circular markers indicate the median, while horizontal and vertical bars indicate the quartile ranges.

The VM performance of the lsb_1-1 model is distinct from the baseline, compared to the ew_gradient model. The PD is in some cases poorer (increased negative timing error) and no clear pattern is observed amongst events for SD and HM.

4.3.5 Discussion: correction procedures

Overall, the main improvement of error weighting comparative to the baseline CVG. Whether or not the prediction envelope captures observed peaks is very important for EWS, especially if the envelopes are being used to generate a flood risk estimate (i.e., the likelihood of the forecasted water level exceeding a certain level). Typically, the EW models have improved timing yet poorer NSE for both the continuous and event-based calculations, which illustrates the shortcomings of measures such as NSE; the model with the strongest NSE value does not correspond to the most useful model in the context of EWS. EW using on the flow gradient produced the largest improvement to timing, ranging from approximately 0 – 2 timesteps (0 – 12 hrs) across the different VMs and events. Despite the improvement in timing caused by gradient weighting, the model performs poorly in every other regard, such as exhibiting inaccurate performance during periods of steady (low) flow and overprediction of high flows. As such it is inadvisable to use gradient weighting as a standalone model without modification such as: increasing the minimum weighting of low flows, taking the gradient over more than a single timestep, or using the gradient weighted model in an ensemble of other models. The other weighting schemes, such as ew_information, were found to be more well-rounded, as the weighting is more evenly spread across different flow values, while still emphasizing infrequent, extreme values thus producing an improvement in timing. Error weighting is encouraged as a very simple method for reducing error associated with specific flow ranges, such as large hydrological events. Future research may consider more sophisticated error weighting schemes may produce greater improvements to timing, such as adaptively weighting error based on localized timing error.

The LSB models typically have worse coverage and narrower precision. LSB typically improved the model timing, albeit, to a lesser degree than models such as ew_gradient. Ultimately, it is not recommended as a means of improving prediction timing – at least without considerable modification. There is a negligible difference in mean ensemble performance while comparing the models that use a learning rate of 0.1 compared to 1, however, the models with a learning rate of 1 were found to be more distinct compared to the baseline. Similarly, the models that have 1 boosting iteration do not perform too distinctly compared to those with 5 boosts and therefore, boosting more than a single iteration is judged to not be worth the additional computational time. Lastly, the LSB models typically outperformed the baseline and EW models based on standard measures; therefore, LSB may be useful for improving performance measures such as NSE, especially if the NSE of the unboosted (baseline) model is poor.

4.3.6 Discussion: visual measures

The following section highlights the strengths and weaknesses of the three VMs used in this research. The main advantage of implementing VMs for flood forecasting is to provide decision-makers with a physically-based understanding of model behaviour, which measures such as NSE do not provide. For example, consider a model with a 4-timestep lead that consistently produces a timing error of two timesteps; a decision-maker interpreting this model will understand that high flows may occur sooner than the model suggests, allowing for flood management precautions to be taken sooner. Otherwise stated, timing error can provide a better estimate of the actual lead time, compared to the lead time for which the model was calibrated. Despite their utility in the context of EWS, VMs are challenging to implement successfully and may not consistently characterize amplitude and timing error.

SD and HM are sensitive to the start and end points identified for each hydrological event. Using a low threshold to distinguish hydrological events may produce SD and HM error quantities that are too low, as timing error is difficult to identify during typical steady flows. Issues associated with event identification may be remedied by using more sophisticated approach to identifying event start and end points, or the VM methods may be adapted for use with a continuous dataset.

While PD is a simple VM, its main weakness is that the identification of the peak flow is subjective, especially for watersheds such as the Bow where high flows are sustained across several timesteps (i.e., once the peak occurs, the flowrate plateaus for several timesteps before decreasing). This research considered the peak flow as the maximum flow during each hydrological event. In cases where the flow level increases and plateaus, the maximum may occur at the rightmost point of the plateau, which is problematic because the leftmost point of the plateau, directly following the rising limb, is more important to predict accurately. Moreover, this may cause an underestimation of timing error, if the predicted high flow plateau has its peak to the right, directly after the rising limb, causing the peaks to be aligned temporally and producing a timing error of 0, when in fact, a timing error is present (e.g., event 3 in Figure 4-6). A more prudent approach to peak flow identification may be consider the flow gradient to the left of the peak, such that if the high-water level plateaus for some amount of time, the leftmost high flow will be taken as the peak.

Similar to PD, SD has difficulty with cases where there is a high flow plateau. Since SD creates segment pairs based on peaks and valleys, it is difficult to calculate an accurate match during periods where the flow is high and steady. Modifying the SD classification scheme to include an additional category for ‘constant flow’, to distinguish these sections from the rising and falling limbs, may improve the poor segmentation near high flow plateaus. HM is not affected by the challenges related to high flow plateaus, however, the effects of the timing error being ‘averaged out’ by low timing error rays is more pronounced in this method. Also, the method is insensitive to erratic predictions (e.g., those made by the ew_gradient model) due to its ability to skip predicted point(s).

Ultimately, VMs provide a broad sense for timing and amplitude error, but each have their own disadvantages. While they are useful for identifying model deficiencies that may be overlooked by standard performance measures such as NSE on the entire dataset; however, the same conclusions could be drawn by applying standard performance measures on an event basis. Ultimately, VMs require additional research related to topics such as sensitivity and data-driven calibration before they are recommended as a useful tool for evaluating hydrological models.

4.3.7 Impact of error weighting on input usefulness

The impact of error weighting can also be assessed using model-based IVS, specifically the CNPS method described in 3.2.2.4. The CNPS values (only values greater than 0.7 are shown) for the ew_gradient models are plotted against those for the baseline models in Figure 4-9 below. Points above the 1:1 line are input variables that are estimated to be more useful in the ew_gradient model compared to the baseline, whereas points below the 1:1 line correspond to inputs that were more useful in the baseline model. For the sake of readability and due to the shear number of inputs, individual inputs are not labelled; rather, inputs are grouped (by colour) by monitoring station. There is not a drastic difference in input usefulness; however, the ew_gradient model tends to utilize precipitation and temperature more consistently compared to the baseline model. Increases use of exogenous inputs, such as precipitation and temperature, is favourable as it suggests the model has less reliance on autoregressive inputs. Overreliance on autoregressive inputs is attributable to timing error, as discussed in 4.1.2.1.

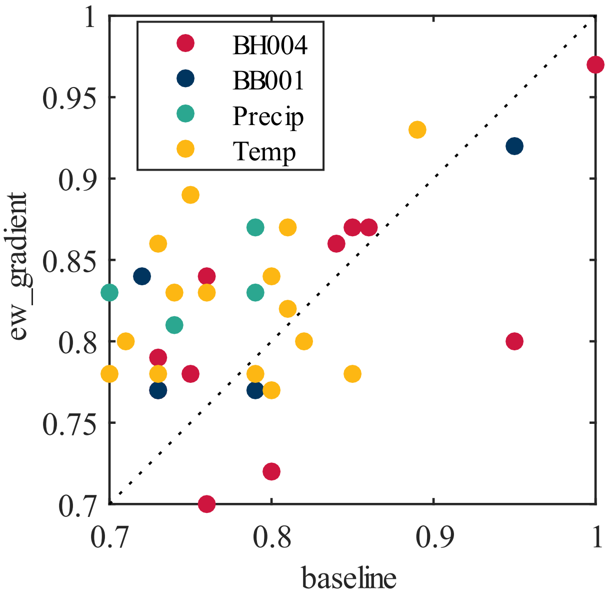

Figure 4-9: Comparison of CNPS values (≥ 0.7) for baseline and ew_gradient models, for input variables at the downstream (red), upstream (blue), precipitation (green), and temperature (yellow) monitoring stations.

4.4 Conclusions and recommendations

This research evaluated two correction procedures for reducing timing error: EW and LSB. While the methods do not directly minimize timing error, they both prioritize flow values that are attributed to having high timing error. VMs, a special class of performance measures that calculate event-based error in terms of amplitude and timing are used to assess the impact of each correction procedure.

This research considered four different EW schemes and four different LSB hyperparameter combinations, along with 3 distinct VMs. Collectively, the various corrected models did not exhibit a considerable change in performance based on NSE and RMSE calculated for the entire dataset. Generally, the error weighted models had a lower NSE and the boosted models had a higher NSE, compared to the baseline. The correction procedures typically improved NSET, however, the improvements were not observed for the test dataset indicating that the models may be poorly generalized. The EW model ensembles showed much better CVG and worse PRC compared to the baseline, whereas the opposite was true for the LSB models.

The ew_gradient and lsb_1-1 models were identified as having the most distinct performance compared to the baseline and were analyzed on an event-basis for 5 events using both standard and visual measures. This analysis revealed more realistic NSE and RMSE values, as the calculation is not biased by the presence of seasonality and the abundance of low, steady flows. The VMs indicate that the ew_gradient model has better timing, and a higher tendency to generate early or over-predictions compared to the baseline, whereas the lsb_1-1 model showed little change relative to the baseline. The event-based performance assessment raises the question of whether or not VMs are necessary for identifying model characteristics such as timing error, or whether standard measures such as NSE and NSET, if used correctly, may be sufficient.

VMs require additional research before they are a recommended tool, as they were found to be relatively insensitive compared to standard performance measures. Also, it is entirely possible that the challenges related to timing error in this watershed are attributable to the lack of exogenous, upstream data (one upstream station is used in this research). Future research may compare these results with other watersheds with different characteristics and data availability.

CHAPTER 5. CONCLUSION

Floods are the most costly and frequent natural disaster in Canada. Flood damage is mitigated by providing advanced warnings of flood events, which rely on models to forecast flood conditions. Data-driven models are increasingly being used for flood forecasting applications, due to being simple to develop, having low data requirements, and producing relatively accurate forecasts. However, there are many challenges associated with developing and interpreting such models. The research objectives outlined in section 1.5 are aimed at improving IVS and peak flow performance for flow forecasting models; both research topics have the ultimate goal of improving flood warning systems, hence mitigating flood damage.

Chapter 3 provides a comprehensive comparison of four IVS methods for two distinct watersheds. Two model-based IVS methods are developed. Notably, IO is improved by using a quantitative, complexity-based selection measure for assessing the effects of input omission without retraining. CNPS is improved by refining the selection criterion, placing greater emphasis on consistent input behaviour across the model ensemble; these developments to IVS methods achieve objective 1.5.1.i. These model-based methods were compared with two model-free methods, PC and PMI. The comparison between four distinct IVS methods achieves objective 1.5.1.ii. The comparison determined that model-based IVS methods were found to produce the best performing models, with CNPS being the most consistent IVS method. The best input reduced models typically exhibit marginally better performance compared to the models with no input reduction and converge more consistently, indicated by the narrower ensemble performance spread. As per objective 1.5.1.iii, this chapter evaluated whether termination criteria are an effective means for choosing the number of model inputs; in several cases, results indicate that models using an arbitrary, predefined number of inputs, often outperform those that use termination criteria. It is proposed that IVS be used to rank input usefulness but to abandon the use of termination criteria to determine the number of inputs. Instead, the number of inputs may be determined in the same manner as other ANN hyperparameters, using a forward addition grid-search, where inputs are added in the order determined by the IVS ranking. This research has direct implications on flow forecasting model applications; the benefits of implementing methods described in Chapter 3 include lower computational demands model calibration and lower data requirements. During this research, it was found that many of the Bow River model predictions were often delayed when compared to the observed values, despite these models having very strong NSE performance. The timing error motivated the research in the Chapter 4, which is focused on assessing and improving model timing, particularly during high flows.

Chapter 4 proposes two different correction procedures for peak flow timing, as per objective 1.5.2.i. Since standard error measures such as NSE do not adequately characterize peak flow timing, VMs are used to quantify the impacts of each correction procedure, addressing objective 1.5.2.ii. The two correction procedures include weighting the cost function during ANN calibration and least-squares boosting. To attain objective 1.5.2.iii, performance is assessed using continuous standard measures, event-based standard measures, and event-based VMs. Based on standard performance measures, the most notable improvement caused by error weighting was the coverage, which represents how well the observations are captured within the uncertainty envelope. Boosting produced poorer coverage and higher precision, in other words a narrower uncertainty envelope that captured fewer observations. Both correction procedures typically produced marginal improvements in timing based on the VMs, with the gradient error weighted models exhibiting the strongest improvement of upwards of 6 to 12 hours. However, this improvement comes at the cost of model performance during low flows, and over-predicting high flows. Subsequently, the gradient weighted model is not recommended as a standalone model, but it may be possible to modify the weighting scheme to improve overall performance, or the model may be used in an ensemble framework, alongside different types of models. Ultimately, error weighting is recommended as a simple method for improving the coverage of the uncertainty envelope and generating slight improvements in peak timing.

Collectively, this work contributes to a more complete understanding of applications of data-driven flow forecasting models, by studying methods for improving performance, lowering cost, interpreting performance, and building trust in model reliability. Severe 2019 spring flooding in Canada renewed calls for improved flood forecasting models nationwide, especially in regions with no existing flood warning systems (Brian Hill 2019). DDMs such as those contained throughout this research provide well suited tool for deploying warning systems, particularly in regions with limited data and funding available for developing more elaborate models. While this thesis contributes to a better understanding of DDMs for flow forecasting applications, many new research questions arose throughout this work, which are discussed in the following section.

5.1 Opportunities for future research

The above research outlined several opportunities for future research related to flow forecasting model development which are summarised below. The first opportunity is to explore the coupled optimization of the number of input and hidden nodes for ANNs. In this research, a predefined number of hidden neurons was used for all the models, regardless of the number of inputs. This is because performing a grid-search, which is the typical method used to optimize the hidden layer size, too computationally expensive to perform for every unique selected input set. A major outcome of the IVS research is the recommendation that the number of inputs be determined using a forward addition approach. Since the number of nodes in the input and hidden layers are interrelated properties, the optimum ANN architecture is achieved by optimizing these values concurrently, rather than one after the other.

Next, since error weighting produced the most pronounced improvement to model performance, other weighting schemes could be evaluated. Examples of other weighting include other variants of gradient weighting and adaptive weighting based on a localized timing error. Another opportunity for future research is a holistic review of VMs. This thesis research compared three VMs; however, there are other approaches for distinguishing between timing and amplitude error, such as wavelet-based methods. Another potential improvement to VMs would be to develop a calibration procedure, where the VM parameters are calibrated using synthetic errors. A data-driven approach would eliminate the subjective tuning of each VM, making them more reliable.

REFERENCES

Abrahart, R.J., Anctil, F., Coulibaly, P., Dawson, C.W., Mount, N.J., See, L.M., Shamseldin, A.Y., Solomatine, D.P., Toth, E., and Wilby, R.L. 2012. Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. Progress in Physical Geography, **36**(4): 480–513. doi:10.1177/0309133312444943.

Abrahart, R.J., Heppenstall, A.J., and See, L.M. 2010. Timing error correction procedure applied to neural network rainfall-runoff modelling. Hydrological Sciences-Journal-des Sciences Hydrologiques, (3): 52. doi:10.1623/hysj.52.3.414.

Abrahart, R.J., See, L., and Kneale, P.E. 2001. Investigating the role of saliency analysis with a neural network rainfall-runoff model. Computers and Geosciences, **27**(8): 921–928. doi:10.1016/S0098-3004(00)00131-X.

Abrahart, R.J., See, L.M., and Solomatine, D.P. 2008. Practical Hydroinformatics. Available from www.springer.com/series/6689.

Almeida, A.M., Castel-Branco, M.M., and Falcao, A.C. 2002. Linear regression for calibration lines revisited: weighting schemes for bioanalytical methods. *In* Journal of Chromatography B. Available from www.elsevier.com/locate/chromb [accessed 11 July 2019].

Anctil, F., and Lauzon, N. 2004. Generalisation for neural networks through data sampling and training procedures, with applications to streamflow predictions. Hydrology and Earth System Sciences, **8**(5): 940–958. doi:10.5194/hess-8-940-2004.

Andy Clark. 2013. $33M for Calgary flood mitigation projects announced | CBC News. Available from https://www.cbc.ca/news/canada/calgary/flood-mitigation-calgary-ottawa-bonnybrook-sunnyside-stormwater-wastewater-funding-1.3845340 [accessed 17 August 2019].

ASCE. 2000a. Artificial Neural Networks in Hydrology. I: Preliminary Concepts. *In* Journal of Hydrologic Engineering. doi:10.1061/(ASCE)1084-0699(2000)5:2(115).

ASCE. 2000b. Artificial Neural Networks in Hydrology. II: Hydrological Applications. *In* Journal of Hydrologic Engineering. doi:10.5121/ijsc.2012.3203.

Barzegar, R., Asghari Moghaddam, A., Adamowski, J., and Ozga-Zielinski, B. 2018. Multi-step water quality forecasting using a boosting ensemble multi-wavelet extreme learning machine model. Stochastic Environmental Research and Risk Assessment, **32**(3): 799–813. Springer Berlin Heidelberg. doi:10.1007/s00477-017-1394-z.

Belayneh, A., Adamowski, J., Khalil, B., and Quilty, J. 2016. Coupling machine learning methods with wavelet transforms and the bootstrap and boosting ensemble approaches for drought prediction. Atmospheric Research, **172**–**173**: 37–47. doi:10.1016/j.atmosres.2015.12.017.

Bennett, N.D., Croke, B.F.W.F., Guariso, G., Guillaume, J.H.A.H., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., and Land, C. 2013. Characterising performance of environmental models. Environmental Modelling and Software, **40**: 1–20. doi:10.1016/j.envsoft.2012.09.011.

Blaszczak, J.R., Delesantro, J.M., Urban, D.L., Doyle, M.W., and Bernhardt, E.S. 2019. Scoured or suffocated: Urban stream ecosystems oscillate between hydrologic and dissolved oxygen extremes. Limnology and Oceanography, **64**(3): 877–894. John Wiley & Sons, Ltd. doi:10.1002/lno.11081.

Bowden, G.J., Dandy, G.C., and Maier, H.R. 2005a. Input determination for neural network models in water resources applications. Part 1 - Background and methodology. Journal of Hydrology, **301**(1–4): 75–92. doi:10.1016/j.jhydrol.2004.06.021.

Bowden, G.J., Maier, H.R., and Dandy, G.C. 2005b. Input determination for neural network models in water resources applications. Part 2. Case study: Forecasting salinity in a river. Journal of Hydrology, **301**(1–4): 93–107. Elsevier. doi:10.1016/j.jhydrol.2004.06.020.

Brian Hill. 2019. Canada is the only G7 country without a national flood forecasting system. Experts say there's a cost to that - National | Globalnews.ca. Available from https://globalnews.ca/news/5221630/canada-is-the-only-g7-country-without-a-national-flood-forecasting-system-experts-say-theres-a-cost-to-that/ [accessed 19 August 2019].

Conway, A.J., Macpherson, K.P., and Brown, J.C. 1998. Delayed time series predictions with neural networks. *In* Neurocomputing. Available from https://ac-els-cdn-com.ezproxy.library.yorku.ca/S0925231297000702/1-s2.0-S0925231297000702-main.pdf?_tid=fbb40819-89af-490a-be04-1f5b8ddeacbc&acdnat=1547491241_63c968e2f7f5da08e400dd39926809f9 [accessed 14 January 2019].

Crochemore, L., Perrin, C., Andréassian, V., Ehret, U., Seibert, S.P., Grimaldi, S., Gupta, H., and Paturel, J.-E. 2014. Comparing expert judgement and numerical criteria for hydrograph evaluation. Hydrological Sciences Journal, **60**(3): 402–423.

doi:10.1080/02626667.2014.903331.

D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, and T. L. Veith. 2007. Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. Transactions of the ASABE, **50**(3): 885–900. American Society of Agricultural and Biological Engineers. doi:10.13031/2013.23153.

Dawson, C.W.W., and Wilby, R.L.L. 2001. Hydrological modelling using artificial neural networks. Progress in Physical Geography, **25**(1): 80–108. doi:10.1191/030913301674775671.

Duncan, A. 2014. The Analysis and Application of Artificial Neural Networks for Early Warning Systems in Hydrology and the Environment. University of Exeter,. Available from http://files/78/Duncan_2014_The Analysis and Application of Artificial Neural Networks for Early Warning.pdf.

Ehret, U., and Zehe, E. 2011. Series distance - An intuitive metric to quantify hydrograph similarity in terms of occurrence, amplitude and timing of hydrological events. Hydrology and Earth System Sciences, **15**(3): 877–896. doi:10.5194/hess-15-877-2011.

Erdal, H.I., and Karakurt, O. 2013. Advancing monthly streamflow prediction accuracy of CART models using ensemble learning paradigms. Journal of Hydrology, **477**: 119–128. Elsevier. doi:10.1016/j.jhydrol.2012.11.015.

Ewen, J. 2011. Hydrograph matching method for measuring model performance. Journal of Hydrology, **408**(1–2): 178–187. doi:10.1016/j.jhydrol.2011.07.038.

Friedman, J.H. 2002. Stochastic gradient boosting. Computational Statistics and Data Analysis, **38**(4): 367–378. doi:10.1016/S0167-9473(01)00065-2.

Gupta, H. V., Kling, H., Yilmaz, K.K., and Martinez, G.F. 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of Hydrology, **377**(1–2): 80–91. doi:10.1016/j.jhydrol.2009.08.003.

He, J., Valeo, C., Chu, A., and Neumann, N.F. 2011. Prediction of event-based stormwater runoff quantity and quality by ANNs developed using PMI-based input selection. Journal of Hydrology, **400**(1–2): 10–23. doi:10.1016/j.jhydrol.2011.01.024.

van Heijst, D., Potharst, R., and van Wezel, M. 2008. A support system for predicting eBay end prices. Decision Support Systems, **44**(4): 970–982. North-Holland. doi:10.1016/j.dss.2007.11.004.

Islam, A.A.S., and Islam, A.A.S. 2010. Improving flood forecasting in Bangladesh using an artificial neural network. Journal of Hydroinformatics, **12**(3): 351. doi:10.2166/hydro.2009.085.

J.~H.~Friedman. 2000. Greedy Function Approximation: A Gradient Boosting Machine. Annals of Statistics, **29**(5): 1189–1232. Available from https://www-jstor-org.ezproxy.library.yorku.ca/stable/pdf/2699986.pdf?refreqid=excelsior%3Aaa503515200994673854453d136913c3 [accessed 11 July 2019].

Jongman, B. 2018, December 29. Effective adaptation to rising flood risk. Nature Publishing Group. doi:10.1038/s41467-018-04396-1.

Kaastra, I., and Boyd, M. 1996. Designing a neural network for forecasting financial and economic time series. Neurocomputing, **10**(3): 215–236. doi:10.1016/0925-2312(95)00039-9.

Kauffeldt, A., Wetterhall, F., Pappenberger, F., Salamon, P., and Thielen, J. 2016. Technical review of large-scale hydrological models for implementation in operational flood forecasting schemes on continental level. Environmental Modelling & Software, **75**: 68–76. Elsevier. doi:10.1016/J.ENVSOFT.2015.09.009.

Khan, U., and Valeo, C. 2017. Optimising Fuzzy Neural Network Architecture for Dissolved Oxygen Prediction and Risk Analysis. Water, **9**(6): 381. doi:10.3390/w9060381.

Khan, U.T., He, J., and Valeo, C. 2018. River flood prediction using fuzzy neural networks: an investigation on automated network architecture. Water Science and Technology, **2017**(1): 238–247. doi:10.2166/wst.2018.107.

Khan, U.T., and Valeo, C. 2016a. Dissolved oxygen prediction using a possibility theory based fuzzy neural network. Hydrology and Earth System Sciences, **20**(6): 2267–2293. doi:10.5194/hess-20-2267-2016.

Khan, U.T., and Valeo, C. 2016b. Short-term peak flow rate prediction and flood risk assessment using fuzzy linear regression. Journal of Environmental Informatics, **28**(2): 71–89. doi:10.3808/jei.201600345.

Kitanidis, P.K., and Bras, R.L. 1980. Real-time forecasting with a conceptual hydrologic model: 2. Applications and results. Water Resources Research, **16**(6): 1034–1044. Wiley-Blackwell. doi:10.1029/WR016i006p01034.

Latecki, L.J., Megalooikonomou, V., Wang, Q., Lakaemper, R., Ratanamahatana, C.A., and Keogh, E. 2005. Elastic partial matching of time series. *In* Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). pp. 577–584. doi:10.1007/11564126_60.

Laureano-Rosario, A., Duncan, A., Mendez-Lazaro, P., Garcia-Rejon, J., Gomez-Carro, S., Farfan-Ale, J., Savic, D., and Muller-Karger, F. 2018. Application of Artificial Neural Networks for Dengue Fever Outbreak Predictions in the Northwest Coast of Yucatan, Mexico and San Juan, Puerto Rico. Tropical Medicine and Infectious Disease, **3**(1): 5. Multidisciplinary Digital Publishing Institute. doi:10.3390/tropicalmed3010005.

Li, S., Ma, K., Jin, Z., and Zhu, Y. 2016. A new flood forecasting model based on SVM and boosting learning algorithms. *In* 2016 IEEE Congress on Evolutionary Computation, CEC 2016. pp. 1343–1348. doi:10.1109/CEC.2016.7743944.

Liem Vu. 2013. Gallery: Severe flooding on DVP in Toronto | Globalnews.ca. Available from https://globalnews.ca/news/597736/gallery-severe-flooding-on-torontos-don-valley-parkway/ [accessed 17 August 2019].

Liu, Y., Brown, J., Demargne, J., and Seo, D.J. 2011. A wavelet-based approach to assessing timing errors in hydrologic predictions. Journal of Hydrology, **397**(3–4): 210–224. doi:10.1016/j.jhydrol.2010.11.040.

Maier, H.R., and Dandy, G.C. 2000. Neural networks for the prediction and forecasting of water resources variables: A review of modelling issues and applications. Environmental Modelling and Software, **15**(1): 101–124. doi:10.1016/S1364-8152(99)00007-9.

Maier, H.R., Jain, A., Dandy, G.C., and 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. Environmental Modelling and Software, **25**(8): 891–909. doi:10.1016/j.envsoft.2010.02.003.

Maier, H.R., Razavi, S., Kapelan, Z., Matott, L.S., Kasprzyk, J., and Tolson, B.A. 2018. Introductory overview: Optimization using evolutionary algorithms and other metaheuristics. Environmental Modelling & Software,. Elsevier. doi:10.1016/J.ENVSOFT.2018.11.018.

Manfreda, S., Mita, L., Dal Sasso, S.F., Samela, C., and Mancusi, L. 2018. Exploiting the use of physical information for the calibration of a lumped hydrological model. Hydrological Processes, **32**(10): 1420–1433. John Wiley & Sons, Ltd. doi:10.1002/hyp.11501.

Masters, T. 1993. Practical Neural Network Recipies in C++. *In* Practical Neural Network Recipies in C++. Morgan Kaufmann. doi:10.1016/c2009-0-22399-3.

May, R., Dandy, G., and Maier, H. 2011. Review of Input Variable Selection Methods for Artificial Neural Networks. *In* Artificial Neural Networks - Methodological Advances and Biomedical Applications. InTech. doi:10.5772/16004.

May, R.J., Dandy, G.C., Maier, H.R., and Nixon, J.B. 2008a. Application of partial mutual information variable selection to ANN forecasting of water quality in water distribution systems. Environmental Modelling and Software, **23**(10–11): 1289–1299. doi:10.1016/j.envsoft.2008.03.008.

May, R.J., Maier, H.R., Dandy, G.C., and Fernando, T.M.K.G. 2008b. Non-linear variable selection for artificial neural networks using partial mutual information. Environmental

Modelling and Software, **23**(10–11): 1312–1326. doi:10.1016/j.envsoft.2008.03.007.

Meshram, S.G., Ghorbani, M.A., Shamshirband, S., Karimi, V., and Meshram, C. 2018. River flow prediction using hybrid PSOGSA algorithm based on feed-forward neural network. Soft Computing, **21**(15): 1–10. doi:10.1007/s00500-018-3598-7.

Mosavi, A., Ozturk, P., Chau, K.W., Mosavi, A., Ozturk, P., and Chau, K.W. 2018. Flood prediction using machine learning models: Literature review. *In* Water (Switzerland). Multidisciplinary Digital Publishing Institute. doi:10.3390/w10111536.

Nanda, T., Sahoo, B., Beria, H., and Chatterjee, C. 2016. A wavelet-based non-linear autoregressive with exogenous inputs (WNARX) dynamic neural network model for real-time flood forecasting using satellite-based rainfall products. Journal of Hydrology, **539**: 57–73. doi:10.1016/j.jhydrol.2016.05.014.

Nath, R., Rajagopalan, B., and Ryker, R. 1997. Determining the Saliency of Input Variables in Neural Network Classifiers. **24**(8): 767–773. Pergamon. doi:10.1016/S0305-0548(96)00088-3.

National Research Council. 2008. Urban Stormwater Management in the United States. Available from www.nap.edu [accessed 17 August 2019].

Office of the Parlimentary Budget Officer. 2016. Estimate of the Average Annual Cost for Disaster Financial Assistance Arrangements due to Weather Events. Available from www.pbo-dpb.gc.ca [accessed 16 August 2019].

Pauline, K.D., See, L., and Smith, M. a. 2001. Towards defining evaluation measures for neural network forecasting models. *In* Proceedings of the Sixth International Conference on GeoComputation. Queensland. p. 11. Available from http://www.geog.leeds.ac.uk.ezproxy.library.yorku.ca/groups/geocomp/2001/papers/kneale.pdf [accessed 30 January 2019].

Public Safety Canada. (n.d.). Floods. Available from https://www.publicsafety.gc.ca/cnt/mrgnc-mngmnt/ntrl-hzrds/fld-en.aspx [accessed 16 August 2019].

Ridgeway, G. 1999. The State of Boosting. Computing Science and Statistics, **31**: 172–181. Available from https://pdfs.semanticscholar.org/1aac/6453fbb8333ee638b6d8b2bb2aff06c3654b.pdf [accessed 7 May 2019].

Schapire, R.E. 1990. The strength of weak learnability. Machine Learning, **5**(2): 197–227. Kluwer Academic Publishers. doi:10.1007/BF00116037.

Seibert, S.P., Ehret, U., and Zehe, E. 2016. Disentangling timing and amplitude errors in streamflow simulations. Hydrology and Earth System Sciences, **20**(9): 3745–3763. doi:10.5194/hess-20-3745-2016.

Setiono, R., and Liu, H. 1997. Neural-network feature selector. doi:10.1109/72.572104.

Sharma, A. 2000. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 - A strategy for system predictor identification. Journal of Hydrology, **239**(1–4): 232–239. doi:10.1016/S0022-1694(00)00346-2.

Sharma, A., Luk, K.C., Cordery, I., and Lall, U. 2000. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 2 - Predictor identification of quarterly rainfall using ocean-atmosphere information. Journal of Hydrology, **239**(1–4): 240–248. doi:10.1016/S0022-1694(00)00347-4.

Shrestha, R.R., and Nestmann, F. 2009. Physically Based and Data-Driven Models and Propagation of Input Uncertainties in River Flood Prediction. Journal of Hydrologic Engineering, **14**(12): 1309–1319. doi:10.1061/(ASCE)HE.1943-5584.0000123.

Shrubsole, D., Kreutzwiser, R., Mitchell, B., Dickinson, T., and Joy, D. 1993. The history of flood damages in ontario. Canadian Water Resources Journal, **18**(2): 133–143. doi:10.1080/cwrj1802133.

Šindelář, R., and Babuška, R. 2004. Input selection for nonlinear regression models. IEEE Transactions on Fuzzy Systems, **12**(5): 688–696. doi:10.1109/TFUZZ.2004.834810.

Solomatine, D.P., and Ostfeld, A. 2008. Data-driven modelling: some past experiences and new approaches. Journal of Hydroinformatics, **10**(1): 3. doi:10.2166/hydro.2008.015.

Strutz, T. 2015. Data Fitting and Uncertainty - A practical introduction to weighted least squares and beyond. *In* Springer Vieweg. Springer. doi:10.1007/978-3-8348-9813-5.

Talei, A., and Chua, L.H.C. 2012. Influence of lag time on event-based rainfall-runoff modeling using the data driven approach. doi:10.1016/j.jhydrol.2012.03.027.

Thiesen, S., Darscheid, P., and Ehret, U. 2019. Identifying rainfall-runoff events in discharge time series: a data-driven method based on information theory. Hydrology and Earth System Sciences, **23**(2): 1015–1034. doi:10.5194/hess-23-1015-2019.

Tompa, D., Morton, J., and Jernigan, E. 2002. Perceptually based image comparison. *In* Proceedings 2000 International Conference on Image Processing (Cat. No.00CH37101). IEEE. pp. 489–492. doi:10.1109/icip.2000.901002.

Tongal, H., and Booij, M.J. 2018. Simulation and forecasting of streamflows using machine learning models coupled with base flow separation. Journal of Hydrology, **564**: 266–282. doi:10.1016/j.jhydrol.2018.07.004.

Toronto and Region Conservation Authority. 2018. Understand Flood Risk Management. Available from https://trca.ca/conservation/flood-risk-management/understand/ [accessed 28 August 2018].

De Vos, N.J., and Rientjes, T.H.M. 2005. Constraints of ANNs for rainfall-runoff modelling Constraints of artificial neural networks for rainfall-runoff modelling: trade-offs in hydrological state representation and model evaluation Constraints of ANNs for rainfall-runoff modelling. HESSD Earth Syst. Sci. Discuss, **2**(2): 365–415. doi:https://doi.org/10.5194/hess-9-111-2005.

Wijesekara, G.N., Gupta, A., Valeo, C., Hasbani, J.G., Qiao, Y., Delaney, P., and Marceau, D.J. 2012. Assessing the impact of future land-use changes on hydrological processes in the Elbow River watershed in southern Alberta, Canada. Journal of Hydrology, **412–413**: 220–232. doi:10.1016/j.jhydrol.2011.04.018.

Wilby, R.. L. 2006. A Review of Climate Change. Built Environment, **33**(1): 31–45. doi:https://doi.org/10.2148/benv.33.1.31.

Wilby, R.. L. 2007. A Review of Climate Change Impacts on the Built Environment. Built

Environment, **33**(1): 31–45. doi:10.2148/benv.33.1.31.

Yilmaz, K.K., Adler, R.F., Tian, Y., Hong, Y., and Pierce, H.F. 2010. Evaluation of a satellite-based global flood monitoring system. International Journal of Remote Sensing, **31**: 3763–3782. doi:10.1080/01431161.2010.483489.

Yin, X., Zhang, J., and Wang, X. 2004. Sequential injection analysis system for the determination of arsenic by hydride generation atomic absorption spectrometry. *In* Fenxi Huaxue. Springer, Berlin. doi:10.1017/CBO9781107415324.004.

Zappa, M., Fundel, F., and Jaun, S. 2013. A "Peak-Box" approach for supporting interpretation and verification of operational ensemble peak-flow forecasts. Hydrological Processes, **27**(1): 117–131. John Wiley & Sons, Ltd. doi:10.1002/hyp.9521.

APPENDIX A. ADDITIONAL MATERIAL FOR IVS RESEARCH

This appendix contains additional material for the research on IVS contained in Chapter 3, grouped into two sections: pseudocode and additional results.

A-1 IVS pseudocode

The pseudocode for each of the four IVS methods is include below.

Partial Correlation

```
for j = 1:num_inputs                                       Select input with maximum
    $R_j$=R($C_j$;y)                                       squared correlation from
end                                                        candidate set
                                                           For the remaining candidate
$C_s$=max($R_j^2$)                                         inputs
remove $C_s$ from C and add to S
                                                           Use linear estimator to calculate
for s = 2:num_inputs                                       residuals
    u=Y-$\hat{Y}$(S)
    Calculate $AIC_{s-1}$(u,S)                             Calculate AIC from previous
                                                           step
    if $AIC_{s-1}$>$AIC_{s-2}$
        terminate algorithm, remove $C_s$ from selection   If the AIC has decreased,
    end                                                    terminate algorithm

    for j=1:|C|                                            Use linear estimator to calculate
        vj=$C_j$-$\hat{C}_j$(S)                            residuals for each remaining
        $PC_j$=R($v_j$;u)$^2$                              candidate
    end
                                                           Select candidate corresponding
    $C_s$=max($PC_j$)                                      to maximum PC between
    remove $C_s$ from C and add to S                       residuals u and v
end
```

Partial Mutual Information

```
for j=1:|C|
    $MI_j = MI(C_j;y)$
End

$C_s=max(MI_j)$
remove $C_s$ from C and add to S

for s=2:|C|
    $u=Y-\hat{Y}(S)$
    Calculate $AIC_{s-1}(u,S)$

    if $AIC_{s-1}>AIC_{s-2}$
        terminate algorithm, remove $C_s$ from selection
    end

    for j=1:|C|
        $vj=C_j-\hat{C}_j(S)$
        $PMI_j=MI(v_j;u)$
    end

    $C_s=max(PMI_j)$
    remove $C_s$ from $C$ and add to $S$
end
```

Select input with maximum mutual information from candidate set

For the remaining candidate inputs

Use kernel estimator to calculate residuals

Calculate AIC from previous step

If the AIC has decreased, terminate algorithm

Use kernel estimator to calculate residuals for each remaining candidate

Select candidate corresponding to maximum PMI between residuals u and v

Input Omission

```
**for** $i = 1{:}num\_models$
    train $\hat{Y} = net_i(C,Y)$
    calculate AIC
end

**for** $j=1{:}J$
    **for** $i=1{:}n$
        calculate $net_{ij}$
        $\hat{Y}_j = net_i(C,Y \mid C_j \notin C)$
        calculate $AIC_j$
    end
end

$S = C(AIC_j \geq AIC)$
```

Train ANN ensemble with candidate input set C

For each candidate input

For each ANN within ensemble

net_{ij} is net_i with parameters unique to candidate C_j removed

get output for input omission ANNs

Calculate AIC for input omission

Select inputs based on inequality for AIC distributions calculated based on Kolmogorov-Smirnov 2-sample test

Combined Neural Pathway Strength

```
for i = 1:num_models                                  Train ANN ensemble with
    $net_i$ = train(C,Y)                              candidate input set C on target Y
end
                                                      Calculate CNPS values
CNPS = $W_1 W_2$

for j = 1:|C|                                         Calculate critical percentages
    $\alpha_j$ = max(count(CNPS>0),count(CNPS<0))
end
                                                      Select inputs from candidates
S = C($\alpha \geq 0.95$)                             based on P values
```

A-2 Additional results

The following section includes additional results for the IVS research. Graphs indicating the selection criterion for each IVS method for the first 6 and 12 selected inputs are shown for the Bow and Don, respectively. Next, IVS model performance is included for the Bow and the Don for 3- and 6-hour, and 2- and 3-day lead times, respectively.

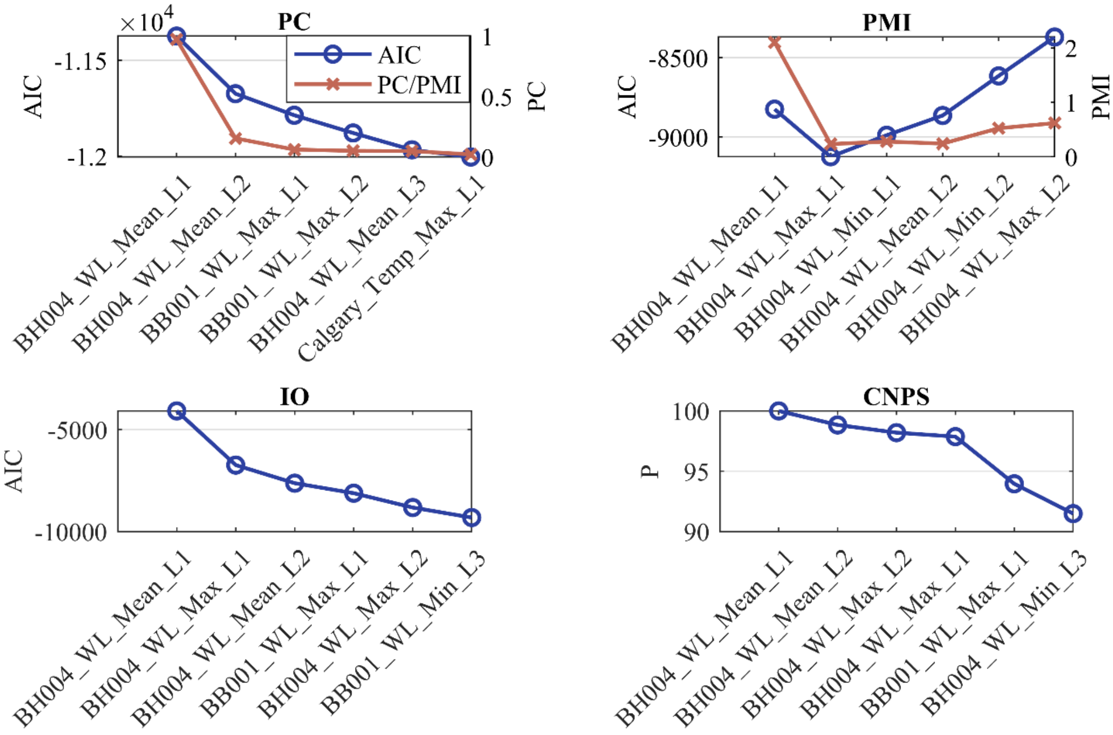

Figure A-1: First 6 input selections for PC, PMI, IO, and CNPS for the Bow River.

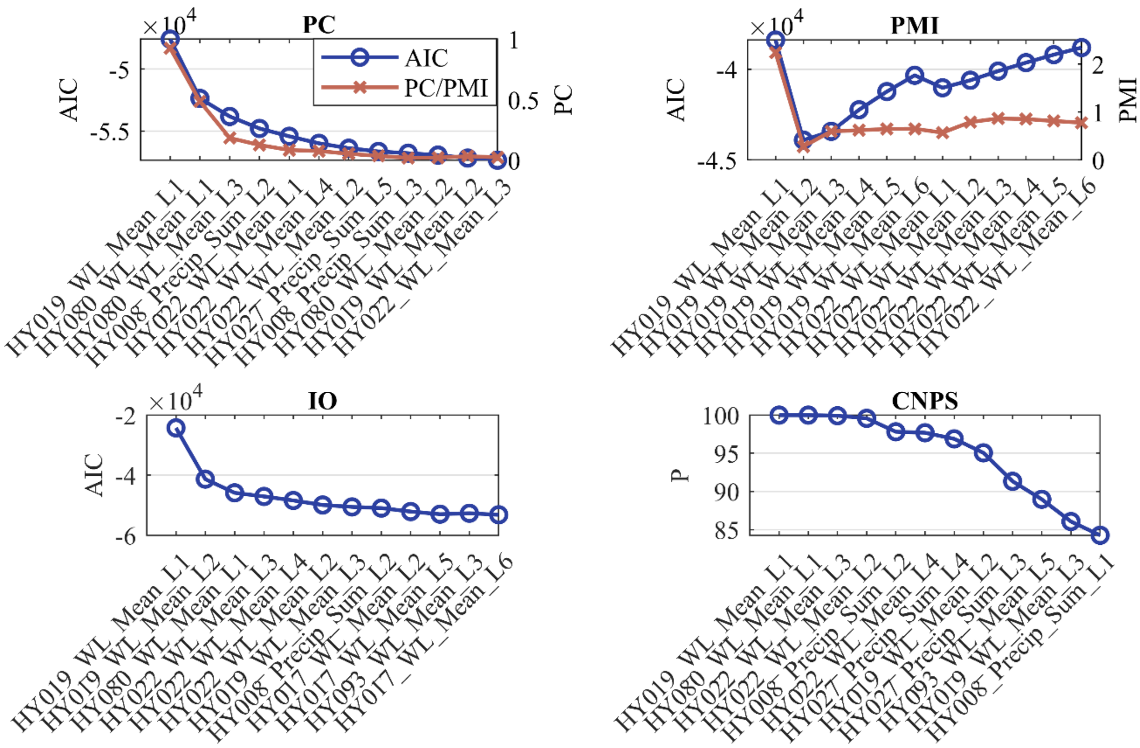

Figure A-2: First 12 input selections for PC, PMI, IO, and CNPS for the Don River.

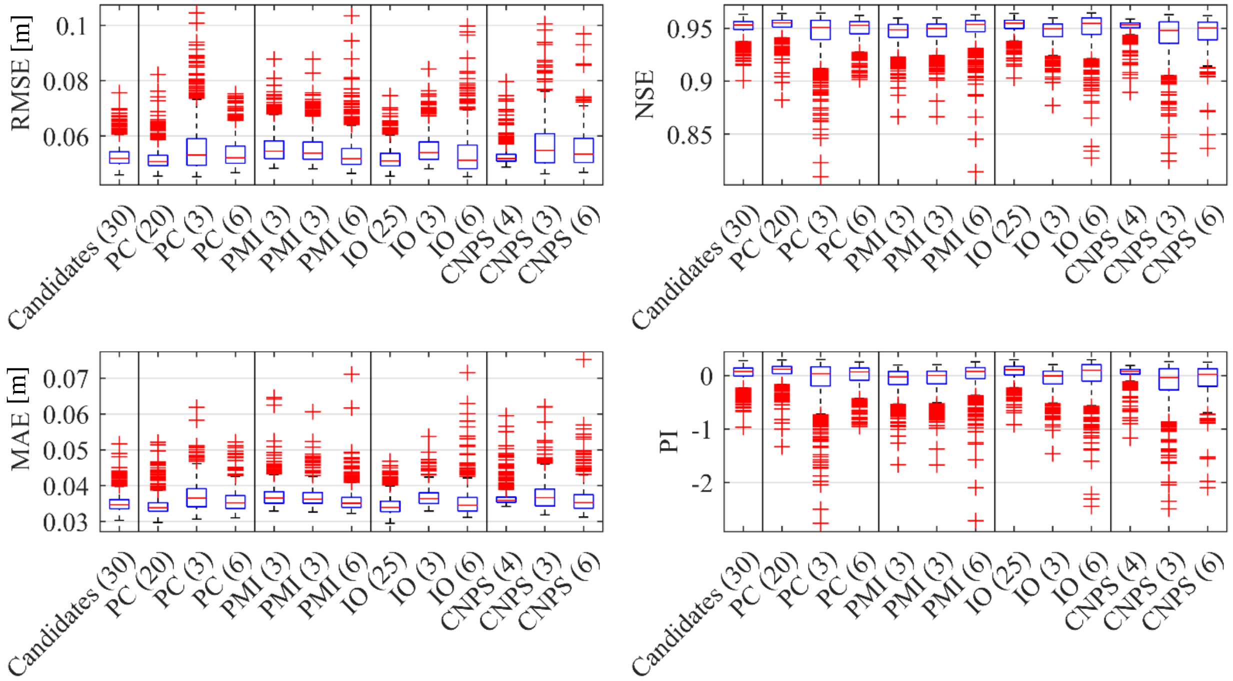

Figure A-3: Comparison of ANN model performance (RMSE, NSE, MAE, and PI) for the Bow River for the 2-day lead time for models that use all candidate inputs (30), termination criteria-based inputs (variable), 10% of all inputs (3), and 20% of all inputs (6) for each of the 4 IVS methods.

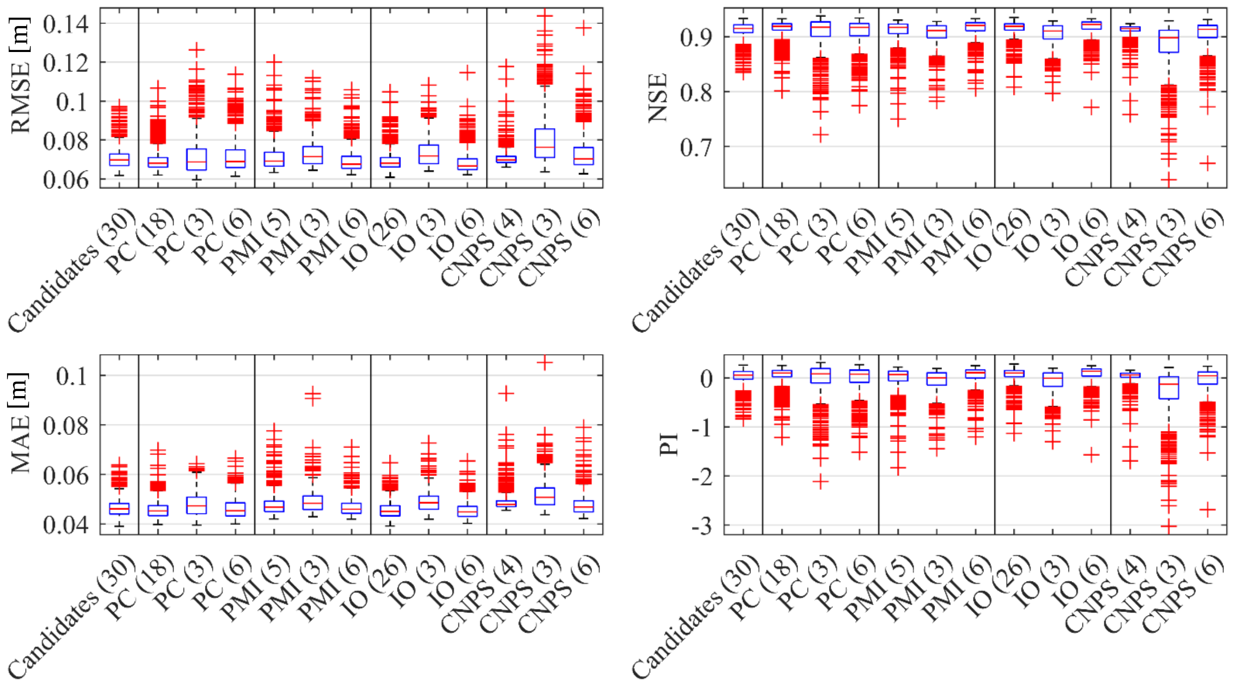

Figure A-4: Comparison of ANN model performance (RMSE, NSE, MAE, and PI) for the Bow River for the 3-day lead time for models that use all candidate inputs (30), termination criteria-based inputs (variable), 10% of all inputs (3), and 20% of all inputs (6) for each of the 4 IVS methods.

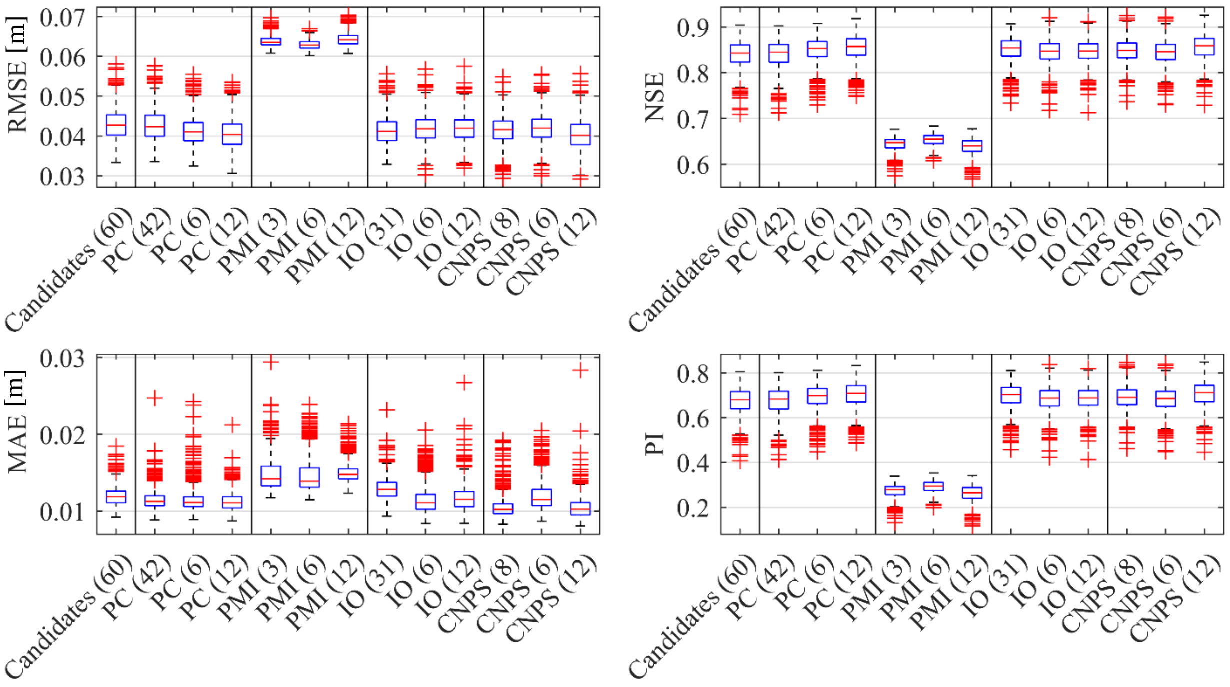

Figure A-5: Comparison of ANN model performance (RMSE, NSE, MAE, and PI) for the Don River for the 3-hour lead time for models that use all candidate inputs (30), termination criteria-based inputs (variable), 10% of all inputs (6), and 20% of all inputs (12) for each of the 4 IVS methods.

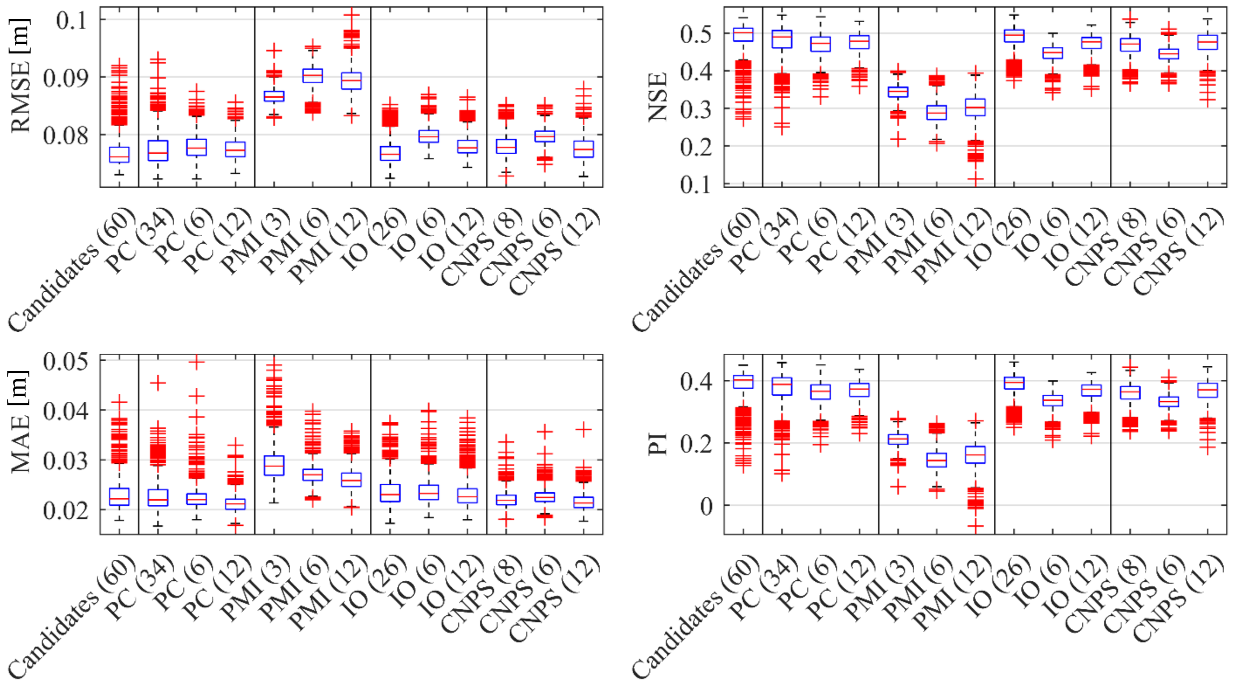

Figure A-6: Comparison of ANN model performance (RMSE, NSE, MAE, and PI) for the Don River for the 6-hour lead time for models that use all candidate inputs (30), termination criteria-based inputs (variable), 10% of all inputs (6), and 20% of all inputs (12) for each of the 4 IVS methods.

APPENDIX B. ADDITIONAL MATERIAL FOR PEAK FLOW PERFORMANCE RESEARCH

This appendix contains additional results for the research on peak flow performance contained in Chapter 4.

B-1 Additional results

Included on the following pages are timeseries and VM performance peakbox plots for all the error weighted and boosted models.

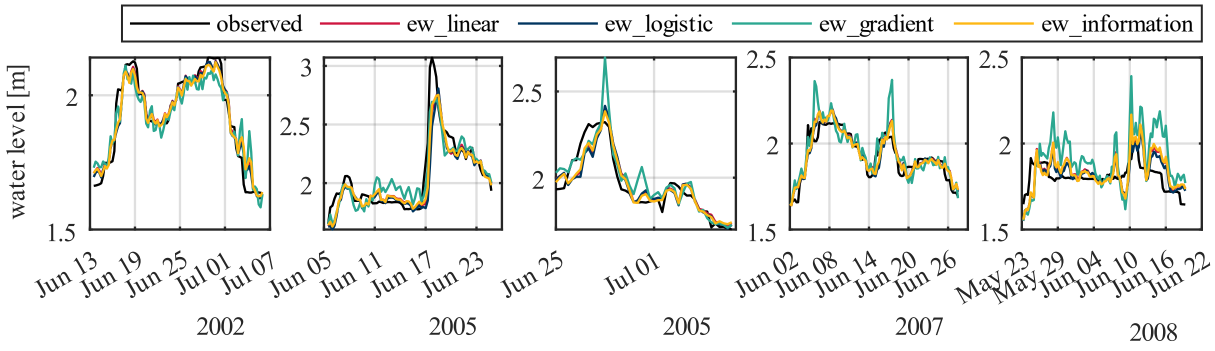

Figure B-1: Observed hydrological events (black) and predicted mean for ew_linear (red), ew_logistic (blue), ew_gradient (green), and ew_information (yellow).

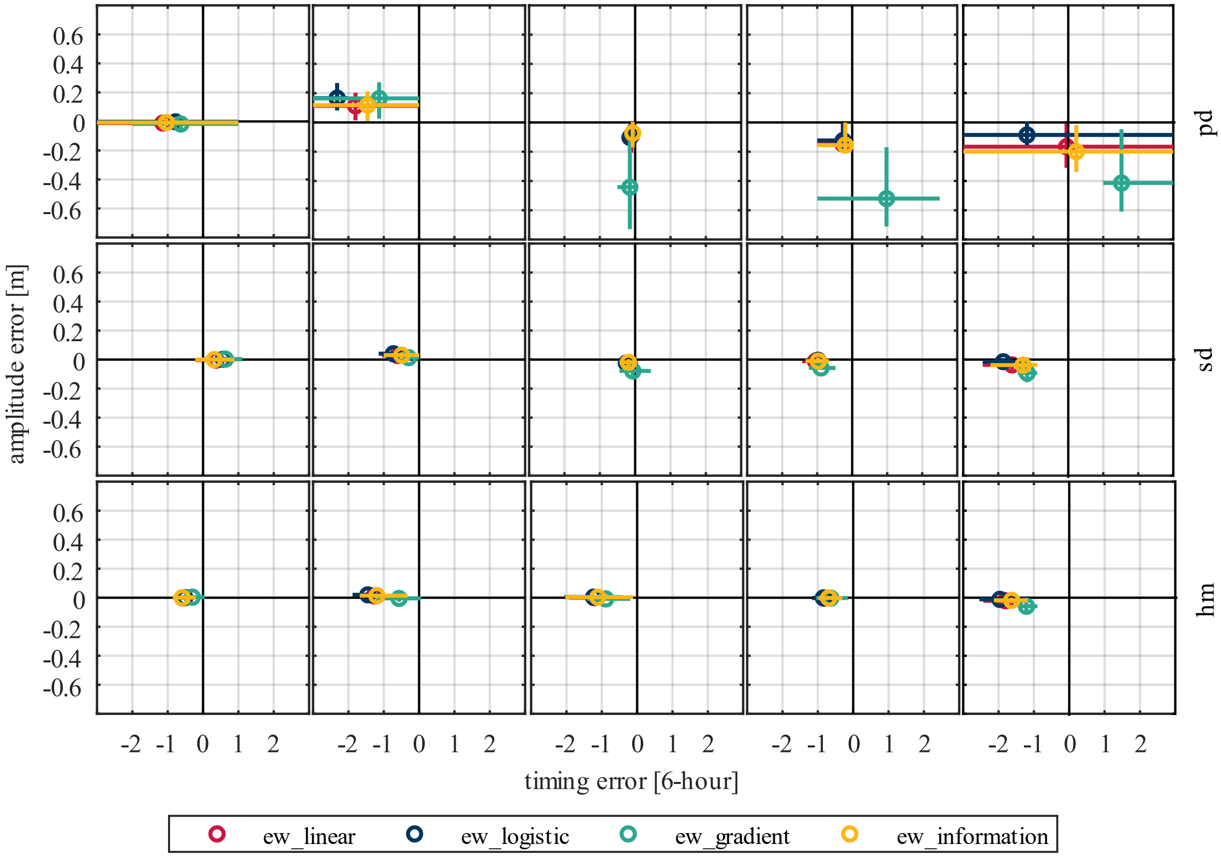

Figure B-2: Performance of visual measures, PD (top row), SD (middle row), HM (bottom row), for ew_linear (red), ew_logistic (blue), ew_gradient (green), and ew_information (yellow).

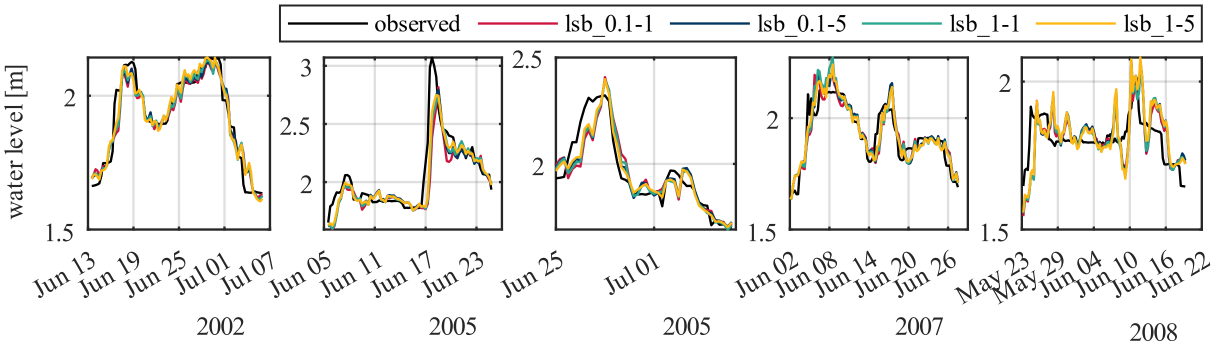

Figure B-3: Observed hydrological events (black) and predicted mean for lsb_0.1-1 (red), lsb_0.1-5 (blue), lsb_1-1 (green), and lsb_1-5 (yellow).

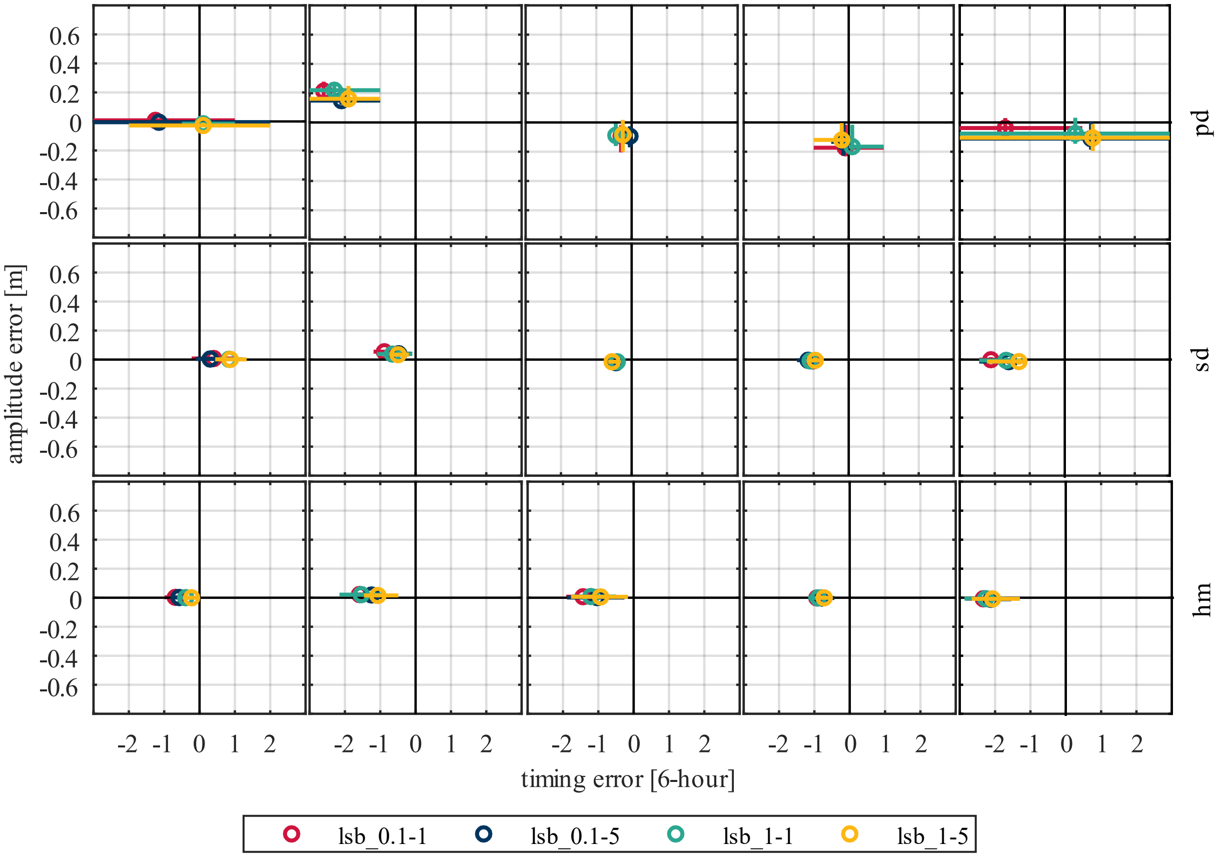

Figure B-4: Performance of visual measures, PD (top row), SD (middle row), HM (bottom row), for lsb_0.1-1 (red), lsb_0.1-5 (blue), lsb_1-1 (green), and lsb_1-5 (yellow).