# Improving hydrological forecasting at multiple lead-times for hydropower reservoir management

by

Behmard SABZIPOUR

MANUSCRIPT-BASED THESIS PRESENTED TO ÉCOLE DE TECHNOLOGIE SUPÉRIEURE IN PARTIAL FULFILLEMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
Ph.D.

MONTRÉAL, JUNE 14, 2023

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

**BOARD OF EXAMINERS**

THIS THESIS HAS BEEN EVALUATED

BY THE FOLLOWING BOARD OF EXAMINERS

Mr. Richard Arsenault, Thesis Supervisor
Department of construction engineering at École de technologie supérieure

Mr. François Brissette, Thesis Co-supervisor
Department of construction engineering at École de technologie supérieure

Mr. Christian Belleau, President of the Board of Examiners
Department of mechanical engineering at École de technologie supérieure

Mrs. Annie Poulin, Member of the jury
Department of construction engineering at École de technologie supérieure

Mr. Alain N. Rousseau, External Evaluator
Centre Eau Terre Environnement, Institut national de la recherche scientifique (INRS)

THIS THESIS WAS PRENSENTED AND DEFENDED

IN THE PRESENCE OF A BOARD OF EXAMINERS AND PUBLIC

MONTRÉAL, MAY 24, 2023

AT ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

## ACKNOWLEDGMENTS

As I reflect on my journey of writing this thesis, a flood of emotions washes over me. Over the years, I faced numerous challenges and obstacles, but I was fortunate to receive invaluable assistance from a multitude of individuals. I am immensely thankful for the unwavering support that allowed me to conduct research and produce this thesis. Without the generous contributions of so many, this achievement would have been impossible.

I wish to express my deep gratitude and appreciation to my Ph.D. supervisor, Dr. Richard Arsenault, for his unwavering support and mentorship throughout my academic journey. Under his guidance, I was able to complete this thesis and achieve my academic goals. Richard has been an exceptional supervisor, always providing valuable advice and scientific insights that helped me overcome research obstacles. His enthusiasm, inspiration, motivation, and patience have encouraged me to make my research a fun and rewarding experience. I am grateful for his clear and concise explanations, which helped me understand complex concepts with ease.

I would like to take a moment to express my sincere appreciation to my co-supervisor, Dr. François Brissette. Francois's humility and unique perspective have made a significant impact on my academic and personal growth, and I am grateful for his unwavering support and guidance.

I would like to express my deepest appreciation to the members of my examining committee, Dr. Annie Poulin, Dr. Christian Belleau. Their invaluable insights and constructive feedback have been instrumental in shaping my work and helping me to reach this point. I am incredibly grateful for their dedication and commitment to ensuring the quality of my research. Without their expert guidance and support, this achievement would not have been possible.

I am filled with gratitude as I reflect on the unwavering support that I have received from both past and present members of the HC3. From the very beginning of my journey, I have been

fortunate to have been met with nothing but help and enjoyable moments of respite that have allowed me to both relax and excel in my studies. Meeting all of you has been a truly unforgettable experience, and I feel immensely fortunate to have found such an incredible group of individuals. It is thanks to the endless contributions of this remarkable community that I have been able to achieve so much. Thank you all.

I would like to thank my co-authors, Magali Troin, Jean-Luc Martel, Frédéric Brunet, Juliane Mai, for their contributions, insights, and suggestions that have significantly improved the quality of this thesis. Their valuable inputs and collaborative efforts have been critical in shaping the outcome of this research. I would also like to extend my appreciation to Magali Troin, Juliane Mai, Jean-Luc Martel for their exceptional assistance in proofreading and editing the papers. Their attention to detail, critical feedback, and suggestions for improvement have been invaluable. I would like to take a moment to acknowledge the immense impact that Philippe Lucas-Picher has had on my academic and personal life. All of the above individual's presence have been a source of tremendous luck, and I am grateful for the opportunity to have known them.

I would like to express my sincere gratitude to the Natural Sciences and Engineering Research Council of Canada (NSERC) and Rio Tinto for the generous financial support, which made this research possible. The contributions provided me with the necessary resources to conduct my research, acquire critical data, and develop innovative solutions. I am immensely grateful for their support, which not only facilitated the completion of this project but also contributed to the advancement of scientific knowledge in my field. The support has been instrumental in helping me achieve my academic and professional goals, and I am honored to have been a recipient of this generosity.

I would like to express my sincere gratitude to the École de technologie supérieure for its unwavering support, particularly during the COVID-19 pandemic. The university's swift and

comprehensive response to the pandemic has been a testament of unconditional commitment to its students, faculty, and staff.

I would like to express my deep gratitude to my family and friends for their unwavering support throughout my academic journey. Your encouragement, love, and understanding have been instrumental in helping me to achieve my goals. Thank you all for being there for me every step of the way.

**Amélioration des prévisions hydrologiques à plusieurs échéances pour la gestion des réservoirs hydroélectriques**

Behmard SABZIPOUR

## RÉSUMÉ

La prévision du débit des cours d'eau est importante pour la gestion des ressources en eau dans des secteurs tels que l'agriculture, l'hydroélectricité, la gestion des sécheresses et la planification de la prévention des inondations urbaines. Notre étude examine les prévisions à court et à long terme afin de créer un cadre pour la prévision des débits qui peut bénéficier à la gestion des ressources en eau et aux secteurs connexes.

Pour améliorer les prévisions de débit jusqu'à dix jours à l'avance, l'étude se concentre d'abord sur l'amélioration des conditions initiales en utilisant un filtre de Kalman d'ensemble comme méthode d'assimilation des données. L'objectif est de réguler les hyperparamètres du filtre de Kalman d'ensemble pour chaque saison afin de produire des prévisions plus précises. Une analyse de sensibilité est menée pour identifier les meilleurs ensembles d'hyperparamètres pour chaque saison, y compris l'incertitude de la température, des précipitations, du débit observé et du contenu en eau de trois variables d'état - zone vadose, zone saturée et manteau neigeux - provenant du modèle CEQUEAU. Les résultats indiquent que l'amélioration des conditions initiales avec le filtre de Kalman d'ensemble produit des prévisions plus habiles jusqu'à un délai de 6 jours. L'incertitude sur la température est particulièrement sensible et varie selon les saisons. La variable d'état de la zone vadose a été identifiée comme la variable d'état la plus importante et la plus sensible, et la mise à jour systématique de toutes les variables d'état n'est peut-être pas nécessaire pour améliorer les prévisions.

Les récentes avancées en matière d'apprentissage automatique permettent d'améliorer les prévisions de débit à court terme. L'une de ces méthodes est le modèle LSTM (Long Short-Term Memory). En général, les réseaux neuronaux apprennent par régression et des relations existent entre les entrées et les sorties. Cependant, les modèles LSTM ont une caractéristique appelée "porte d'oubli", qui leur permet non seulement d'apprendre la relation entre les entrées (par exemple, la température et les précipitations) et la sortie (le débit), mais aussi de saisir les dépendances temporelles dans les données. L'étude visait à comparer les performances du modèle LSTM (Long Short-Term Memory) avec celles des modèles hydrologiques basés sur l'assimilation de données et sur les processus dans la prévision à court terme du débit. Les trois modèles ont été testés en utilisant les mêmes prévisions météorologiques d'ensemble. Le modèle LSTM a démontré une bonne performance dans la prévision du débit, avec une efficacité de Kling-Gupta (KGE) supérieure à 0,88 pour 9 délais. Le modèle LSTM n'a pas incorporé d'assimilation de données, mais il a bénéficié du débit observé jusqu'au dernier jour

avant la prévision. Cela s'explique par le fait que le modèle LSTM a appris et intégré les connaissances des jours précédents tout en émettant des prévisions, de la même manière que l'assimilation de données met à jour les conditions initiales. Les résultats de l'étude ont également montré que le modèle LSTM avait de meilleures performances jusqu'au sixième jour de prévision par rapport aux modèles basés sur l'assimilation de données. Cependant, l'entraînement du modèle LSTM séparément pour chaque délai d'exécution est un processus long et constitue un inconvénient par rapport aux méthodes basées sur l'assimilation de données. Néanmoins, l'étude a démontré le potentiel des techniques d'apprentissage automatique pour améliorer les prévisions de débit.

La prévision du débit pour de longs délais, comme un mois, implique généralement l'utilisation de données météorologiques historiques pour créer des scénarios futurs probables, car les prévisions météorologiques deviennent peu fiables au-delà de ce délai. Dans cette étude, nous avons proposé une nouvelle méthode de prévision du débit basée sur le filtrage des prévisions de débit d'ensemble (ESP), en utilisant un algorithme génétique (GA) pour filtrer les scénarios de prévision. Cette méthode quantifie le potentiel existant dans les données historiques pour chaque bassin. Ce potentiel pourrait être utilisé pour améliorer la précision des prévisions de débit. Nous avons trié les scénarios sélectionnés et non sélectionnés pour trouver les caractéristiques communes entre eux, mais les résultats n'ont pas permis de distinguer les deux groupes. Néanmoins, la méthode GA peut être utilisée comme référence pour de futures études visant à améliorer les prévisions de débit à long terme. Cette méthode peut également être utilisée pour comparer différentes méthodes de prévision en fonction du potentiel démontré par la méthode GA pour une taille spécifique de membres ESP. Par exemple, si une méthode utilise des signaux climatiques à grande échelle pour filtrer les membres de l'ESP, le résultat de la compétence de prévision pourrait être comparé au potentiel des données historiques pour cette taille particulière de membres de l'ESP.

**Mots-clés** : prévision hydrologique ; prévision d'ensemble des débits ; LSTM ; assimilation de données ; incertitude et vérification des prévisions.

# Improving Hydrological Forecasting at Multiple Lead-Times for Hydropower Reservoir Management

Behmard SABZIPOUR

## ABSTRACT

Streamflow forecasting is important for managing water resources in sectors like agriculture, hydropower, drought management, and urban flood prevention planning. Our study examines short and long lead-times to create a framework for streamflow forecasting that can benefit water resource management and related sectors.

To improve streamflow forecasts for up to ten days of lead-time, the study first focuses on improving initial conditions using an ensemble Kalman filter as a data assimilation method. The goal is to regulate the hyperparameters of the ensemble Kalman filter for each season to produce more accurate forecasts. A sensitivity analysis is conducted to identify the best hyperparameter sets for each season, including uncertainty in temperature, precipitation, observed streamflow, and the water content of three state variables - vadose zone, saturated zone, and snowpack - from the CEQUEAU model. Results indicate that improving initial conditions with the ensemble Kalman filter produces more skillful forecasts until a 6-day lead-time. Temperature uncertainty is particularly sensitive and varies across seasons. The vadose zone state variable was identified as the most important and sensitive state variable, and updating all state variables systematically may not be necessary for improving forecast skill.

Recent machine learning advances are improving short-term streamflow forecasting. One such method is the Long Short-Term Memory (LSTM) model. In general, neural networks learn from regression as relationships exist between input-output. However, LSTM models have a feature named ‘forget gate’, which enables them to learn the relationship between inputs (e.g., temperature and precipitation) and output (streamflow), and also to capture temporal dependencies in the data. The study aimed to compare the performance of the Long Short-Term Memory (LSTM) model with data assimilation-based and process-based hydrological models in short-term streamflow forecasting. All three models were tested using the same ensemble weather forecasts. The LSTM model demonstrated good performance in forecasting streamflow, with a Kling-Gupta efficiency (KGE) greater than 0.88 for 9 lead-times. The LSTM model did not incorporate data assimilation, but it benefited from observed streamflow until the last day before the forecast. This is because the LSTM model learned and incorporated knowledge from the previous days while issuing forecasts, similar to how data assimilation updates initial conditions. The study results also showed that the LSTM model had better performance up to day 6 of lead-time compared to the data assimilation-based models. However, training the LSTM model separately for each lead-time is a time-consuming process and is a disadvantage compared to the data assimilation-based methods. Nonetheless, the study

demonstrated the potential of machine learning techniques in improving streamflow forecasting.

The forecasting of streamflow for long lead-times such as a month usually involves the use of historical meteorological data to create probable future scenarios, as meteorological forecasts become unreliable beyond this lead-time. In this study, we proposed a novel method for streamflow forecasting based on ensemble streamflow forecasting (ESP) filtering, using a Genetic Algorithm (GA) to filter forecast scenarios. This method quantifies the potential of historical data for each basin. This potential could be utilized to enhance the accuracy of streamflow forecasts. We sorted the selected and unselected scenarios to find out the common features between them, but the results did not help distinguish between the two groups. Nonetheless, the GA method can be used as a benchmark for future studies to improve long-term streamflow forecasting. This method can also be used to compare different forecast methods based on the potential shown by the GA method for a specific size of ESP members. For instance, if a method uses large-scale climate signals to filter ESP members, the forecast skill result could be compared with the potential of historical data for that particular size of ESP members.

**Keywords**: hydrological forecasting; ensemble streamflow prediction; LSTM; data assimilation; forecast uncertainty and verification.

**TABLE OF CONTENTS**

Page

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBRIEVIATIONS AND ACRONYMS

| | |
|---|---|
| AE | Autoencoder |
| AEnKF | Asynchronous Ensemble Kalman filter |
| AI | Artificial Intelligence |
| AM | April May |
| AMO | Atlantic Multi decadal Oscillation unsmoothed |
| ANN | Artificial Neural Network |
| BC | British Clumbia domain |
| C3S | Climate Change Service |
| CDF | Cumulative distribution function |
| CEMS | Copernicus Emergency Management Service (CEMS) |
| CRPS | Continuous Ranked Probability Score |
| CRPSS | Continuous Ranked Probability Skill Score |
| CT | Contingency Table |
| DA | Data Assimilation |
| DJFM | December January February March |
| ECMWF | European Centre for Medium-Range Weather Forecasts |
| ED | Encoder-Decoder |
| ENSO | El Niño-Southern Oscillation |
| EKF | Extended Kalman Filter |
| EnKF | Ensemble Kalman Filter |
| ESP | Ensemble Streamflow Prediction |

| | |
|---|---|
| GloFAS | Global Flood Awareness system |
| Glo-Sea5-GC2 | Global Seasonal forecast system version 5 using the Global Coupled model configuration 2 |
| GMON | Gamma Ray Monitors |
| KGE | Kling-Gupta Efficiency |
| GA | Genetic Algorithm |
| HMETS | Hydrological Model École de Technologie Supérieure |
| HRU | Hydrological Response Units |
| IFS | Integrated Forecasting System |
| JJA | June July August |
| KF | Kalman Filter |
| LSJ | Lac-Saint-Jean |
| LSTM | Long Short-Term Memory |
| MAE | Mean Absolute Error |
| MARS | Meteorological Archival and Retrieval System |
| MEI | Multivariate El Niño/Southern Oscillation Index |
| NSE | Nash-Sutcliffe Efficiency |
| NWP | Numerical Weather Prediction |
| PDO | Pacific Decadal Oscillation |
| PF | Particle Filter |
| PNA | Pacific-North American |
| REnKF | Recrusive Ensemble Kalman Filter |

| | |
|---|---|
| RNN | Recurrent Neural Netword |
| SAC-SMA | Sacramento Soil Moisture Accounting |
| SON | September October November |
| QC | Quebec domain |
| SOI | Southern Oscillation Index |
| SWE | Snow Water Equivalent |

# INTRODUCTION

Human civilization first occurred along water courses and bodies, and as such, preserving and monitoring water resources have always been high priorities in our societies. Hydrology deals with water movement, from understanding the basic governing physical processes ranging from the very short to the very long terms. To do so, hydrologists have been using hydrological models (HM) to transform meteorological data into river flows. HMs mimic a watershed, considering its physical aspects such as soil moisture, vegetation cover, slope and other properties that affect the flow of water. HMs use meteorological inputs like precipitation and temperature to provide users (or stakeholders) the amount of runoff or streamflow at any desired point(s) in a watershed. However, modeling watershed hydrology is not only done using HMs. In recent years, data-driven (DD) methods have become a promising alternative. In contrast to HMs, which require many types of observational data and a good understanding of the fundamental hydrological processes, DD methods such as Artificial Neural Networks (ANNs) have demonstrated their ability to simulate streamflow without any a priori knowledge of physical processes (Damavandi et al., 2019), although having big data is a challenge.

Accurately simulating the relationship between meteorological data and water movement makes it possible to forecast streamflows from weather forecasts. Reliable and accurate streamflow forecasts are crucial for water-dependent sectors like agriculture, hydropower generation, and flood management (Boucher et al., 2012; Anghileri et al., 2016; Cassagnole et al., 2021). Therefore, forecasting streamflows has been part of the hydrological sciences since the 1970s (Twedt et al., 1977; Day, 1985) and has been improving considerably ever since. Intrinsically, forecasting is an uncertain undertaking and as such, issuing probabilistic forecasts has been the norm for over fourty years to deal with this uncertainty (Krzysztofowicz, 2001).

Furthermore, streamflow forecasting can be performed on multiple horizons (or lead-times), varying from short-term (a few hours) to long-term (a few months). Accordingly, forecasting

methods also vary to account for this variation in forecast lead-time, with specific methods tailored to each narrow lead-time window. This also means that improving the performance of hydrological forecasts will depend on the lead-time of interest. In the case of hydropower generation, multiple lead-times are of interest. First, short-term lead-times help optimize water resources allocation at the level of the turbines in the powerplant. Then, forecasts for up 2 weeks are useful to allocate water between reservoirs and optimize total water allocation per generating station. Finally, longer-term forecasts can be useful to plan power production over longer periods to plan for maintenance, commit to sale-of-energy contracts and optimize long-term reservoir levels (Cuo et al., 2011; Zhao et al., 2012; Monhart et al. 2019; Harrigan et al. 2018).

However, each of these scenarios produce streamflow forecasts that are imperfect in many ways. Hydrological models are simplistic simulators of the hydrological cycle, weather forecasts are highly uncertain, and some processes are simply absent from the modelling tools. (Zappa et al., 2010; Bourgin et al., 2014; Sun et al., 2018; Moradkhani and Sorooshian, 2009; Ajami et al., 2007; Li et al., 2009; Cloke and Pappenberger, 2009). This leads to forecasts that can be biased, unreliable and that do not represent the total uncertainty in an adequate manner. Therefore, there is a need to improve these hydrological forecasts such that water resources managers can make better informed decisions and optimize water usage over the multiple time horizons.

The main objective of this thesis is to improve and further of our understanding of the limitations of both short-term (up to 10 days) and long-term (up to 6 months) hydrological forecasts for a hydropower system. To attain this main objective, the research project is divided into three main research areas: (1) improving long-term hydrological forecasts using hydrological models, (2) improving short-term hydrological model-based forecasts and (3) investigating the applicability of ANNs in short-term forecasting. The latter includes a specific type of ANN called Long Short-Term Memory (LSTM) models that can replace the structure of HMs without specifically being trained to do so. To address these three topics, a hydropower

system in Quebec, Canada, was used as the target watershed as it is owned and operated by the partner organization Rio Tinto for providing energy to their aluminium smelters in the Saguenay-Lac-St-Jean region.

This thesis is divided into the six following chapters. The next chapter presents a literature review to contextualize efforts made by the community in this area of research as of yet. The development and implementation of the methods addressing each of the three secondary objectives are presented in Chapters 3, 4, and 5. respectively, following the thesis organization section in chapter 2. Finally, a general discussion is presented in chapter 6, followed by concluding remarks and recommendations.

# CHAPTER 1

# LITERATURE REVIEW

## 1.1 Importance of hydrological forecast

Hydrological forecasting is not only important for water resources management but also for human security. Agriculture, hydropower, urban planning, defeating natural crises like droughts and floods, are among the sectors that benefit and require hydrological forecasting for efficient operation (Clark & Hay, 2004; Gutiérrez & Dracup, 2001; Rodda, 2011).

### 1.1.1 Hydrological forecasting methods

Hydrological forecasting comes in many different forms. The types and sources of input data (either historical or forecasted weather) as well as the accuracy and uncertainty are highly modulated by the type of forecasting system required and/or implemented. This section presents the multiple facets of hydrological forecasting, from the oldest (and simplest) methods to the state-of-the-art methods.

#### 1.1.1.1 Ensemble Streamflow Prediction

Flood forecasting or in broader view streamflow prediction dates back to the late 1970s and early 1980s. Fears of floods and droughts were the incentives for issuing long-term streamflow forecasts in the United-States (https://hepex.inrae.fr/tracing-the-origins-of-esp/). The first acknowledged method that makes use of hydrological models for forecasting streamflow is the extended streamflow prediction (ESP) method later changed to Ensemble Streamflow Prediction (Day, 1985; Hirsch et al., 1977; Twedt et al., 1977). Forecasters, not privy to the future state of the atmosphere, would use historical weather observations as proxies to actual

possible future weather scenarios (Harrigan et al., 2018). This weather could then be fed to a hydrological model and converted to possible streamflows. The advantage of this method is that a hydrological model would at least have a sensible idea of the current state of the watershed, such that the future weather would have an impact on future streamflows, but this would be conditional on the actual state of the watershed. This was a breakthrough compared to previous methods that simply used historical observations of streamflows as possible future realizations, where the initial state of the watershed could be completely different than the current conditions. Using historical weather, however, does cause problems due to the fact that the current state of the atmosphere is neglected. This means that ESPs could consider a stormy, rainy day as a plausible forecast for the coming days whereas the actual atmospheric conditions could be warm and dry. Therefore, to get the best chance of covering possible outcomes, weather from multiple (ideally many) previous years are used and combined to create an ensemble of possible future streamflows, with each member of the ensemble streamflow forecasts being associated to a single weather forecast (i.e., one historical realization of past weather).

At first, ESP was used for short lead-times for flood forecasting and other important water-related issues. However, with time, it has been slowly replaced by more advanced Numerical Weather Prediction (NWP) systems as discussed in more details in section 1.1.1.2. Since the advent of NWPs meant that the atmosphere could be modeled and provide more precise forecasts than ESPs on shorter lead-times, the latter became obsolete, yet remained useful for longer lead-times only, where the former remain unskillful (Demargne et al., 2009). In essence, ESP is a forecasting method that can provide information on possible future streamflows but is limited in the prediction accuracy due to a lack of sufficient information on the state of the atmosphere at longer lead-times. It is still highly useful today to provide long-term probabilities of streamflow for cases such as hydropower reservoir management, water availability, flood risk and mitigation and other cases where forecasting far in advance could be of use (Alfieri et al., 2014; Monhart et al., 2019; Zhang et al., 2020)

An important factor in ESPs is the reliability and likelihood of historical events being representative of future events. Climate change, changes in the atmospheric patterns at various time scales, teleconnections and natural variability could all be predictors of the future state of the atmosphere and could thus provide more information on how to select the most likely scenarios from the past such that they represent the future weather as best as possible. For example, perhaps some atmospheric conditions existing today could help predict that the following months will be rainier than on average. This might in turn be used to select more rainy members from the historical dataset to condition it to expected future conditions. Thus, issuing forecasts with pre-processing of the ensemble members in advance could improve forecasting skill (Ehsan et al., 2021; Li et al., 2017; Yuan et al., 2015). Improving long-term forecast skill could be attainable by conditioning the selection of ensemble data using climate indices (Najafi et al., 2012), given similarity between climate signals in the period before the forecast with climate signals in the same historical period.

Stastical methods to process streamflow data in such a manner are common in using streamflow ensembles. Conditioning on climate signals (Beckers et al., 2016; Donegan et al., 2021), using the Euclidean distance between ensemble members' streamflow and streamflow of the period before the forecast (Koutsoyiannis et al., 2008; Yao & Georgakakos, 2001), associating weights to ensemble members considering large-scale climate indices (Grantz et al., 2005; Hamlet & Lettenmaier, 1999) and utilizing hydrological persistence as an indicator (Svensson, 2016) are all methods that were reported in the literature to improve the quality of ESP. Furthermore, ESP is not limited to any temporal scale, such that forecasts can be performed on daily (Harrigan et al., 2018; Pagano et al., 2010), weekly (Hapuarachchi et al., 2022), monthly (Fundel et al., 2013), seasonal (Baker et al., 2021) and even inter-annual (Anghileri et al., 2016) time scales.

#### 1.1.1.2 NWP in streamflow forecasting

With the development of weather forecasting models in the past decades, hydrologists gained a new tool to help predict streamflow. NWP systems allow modelling the current state of the atmosphere, and with equations of energy, mass, and momentum conservation, can predict changes in weather across the globe for the near future. These weather forecasts can then be used in hydrological models to predict streamflows. This has the advantage, compared to ESPs, of including the state of the atmosphere in the weather forecasts, improving their accuracy significantly (Cloke & Pappenberger, 2009; Schaake et al., 2006). One important study (Cloke & Pappenberger, 2009) discussed streamflow forecasting using NWP data for flood forecasting in Europe and showed that the NWP-based approach had considerable skill in predicting flood events. NWPs predict weather using a variety of methods, but the two most important for hydrological forecasting are the deterministic (single member) and probabilistic (ensemble) forecasts. NWP forecasts require a lot of computing power and as such, weather forecasting centers will typically provide one "most precise" possible forecast, using high-resolution models and the best possible estimates of the atmospheric variables at the time of forecast. This is the deterministic weather forecast and is typically the most accurate forecast for the operational forecasting model. However, the deterministic forecast lacks a crucial component: it is unable to estimate the uncertainty of the forecast since it is a single value forecast for each lead-time and each location on the forecasting model grid. Therefore, weather forecasting centers also typically issue a probabilistic forecast using a less-refined version of the forecasting model that is much faster to run. This probabilistic forecast can contain a variable number of members, typically between 10 and 50 members each, where each member is generated using different initial conditions of the atmosphere according to the uncertainty of the atmospheric state. Therefore, users can either obtain the most accurate deterministic forecast (but with no uncertainty assessment) and/or the less accurate probabilistic forecast (but allowing for uncertainty to be evaluated and considered). However, in both cases, since NWP models simulate the atmosphere and because the atmosphere is highly chaotic, NWPs are typically only skillful in the 7-to-10-day lead-time range (Stern & Davidson, 2015). After

this period, their forecasts are typically not better than that of the ESP forecasts. Therefore, for short lead-time forecasts such as those used in quick-response flood prediction, using NWP is dominant.

The literature is rife with studies using, evaluating and improving NWP-based hydrological forecasts. Jasper et al. (2002) used a framework based on coupling hydrological models with atmospheric weather forecasts. They utilized observational data from surface monitoring stations and radars, for calibrating and spinning-up their hydrological model, then used NWP predictions to issue forecasts for several extreme flood events. Yucel et al. (2015) used NWP data as inputs of hydrological model to evaluate forecasting skill for flood events as well. The forecast was based on precipitations from the Weather Research and Forecasting (WRF) model and the EUMETSAT Multi-sensor Precipitation Estimates (MPEs), to force a semi-distributed hydraulic-hydrological model, WRF-hydro. They calibrated their model on some of the basins and tested it on other independent basins. Results showed the model was able to predict reasonably well the events regarding the volume of runoff and the behaviour of hydrographs. Shi et al. (2015) produced a webservice to issue flood forecasting on mountain area using NWP inputs. Also, there are plenty of studies which include a two-component framework, including numerical weather forecasts as forcing, and a hydrological model as the translator of forcings to produce corresponding streamflow forecasts. For example, Shi et al. (2015), Collischonn et al. (2005), Ming et al. (2020), Şensoy and Uysal (2012), Smiatek et al. (2012), Zhao et al. (2009) all implemented such methods. There is therefore a solid body of literature supporting this type of forecasting method and it is currently the operational standard for hydrological forecasting on shorter lead-times.

#### 1.1.1.3 Deterministic hydrological forecasting

Deterministic hydrological forecasting refers to any approach that leads to the simulation of a single possible future realization of streamflows. If more than one realization is performed, the forecast will be probabilistic in nature. Deterministic forecasts are thus typically generated

using a single (usually the deterministic forecast but could also be a single member of a probabilistic weather forecast) member of a weather forecast and fed into a hydrological model or other type of rainfall-runoff model. These forecasts, as is the case for the deterministic weather forecasts, are assumed to be the most accurate that the model can produce given the model structure and quality of input data since they are driven by the deterministic weather forecasts. However, they also do not convey any information regarding the uncertainty of the forecast.

The skill of the deterministic hydrological forecast depends on a multitude of factors, including the spatial resolution and scale of the watershed of interest, as well as the lead-time of the desired forecast. For instance, a forecast over the next 24 hours is probably going to be of higher accuracy than another aimed at the 10-day lead-time. This is simply due to the fact that hydrological model errors will compound with time, along with the fact that the input data (the weather forecast) will also diminish in skill at longer lead-times. This is expected; however, it might also provide a false sense of confidence given that there is no uncertainty associated with the deterministic hydrological forecast. The same reasoning can be made for watersheds of different sizes. Indeed, a strong rainfall event might be forecasted to pass beside a small catchment whereas the rainfall actually did land on the catchment. This would lead to a large error caused by the forecast being wrong with respect to the exact location of the event but not its magnitude. Conversely, for a larger basin, the spatial errors average out more evenly and thus lead to more stable forecasting performance (Schaake et al., 2007).

However, as stated, deterministic hydrological forecasts do not provide any uncertainty assessment and must therefore be used with caution (Demargne et al., 2009). They will typically be used for shorter periods (up to a few days) since the uncertainty on very short periods usually remains small and the deterministic forecasts remain of high accuracy. However, as lead-times increase, more and more errors compound leading to the necessity of employing uncertainty assessment measures such as probabilistic forecasts (Fan et al., 2016) (see section 1.1.1.4). This is a major point of interest for applications with desired lead-times

of more than just a few days, such as in hydropower reservoir management. Indeed, uncertainty plays an important role since the decision to turbine more water, or spill excess water, and to generally optimize the process relies heavily on the management of risk, and thus requires a good estimation of uncertainty to quantify this risk efficiently and adequately (Arsenault & Côté, 2019; Ficchì et al., 2016). Schwanenberg et al. (2015) compared three hydrological forecasting methods, including single-member perfect forcasts (observed streamflow), deterministic and probabilistic forecasts as inputs to the stochastic optimization of a hydropower reservoir's operation during flood events. They found that probabilistic forecasts were as skillful as perfect (observation-based) forecasts, both beating the deterministic forecast, showing the importance of uncertainty assessment in hydrological forecasting for specific applications. Boucher et al. (2011) found that probabilistic forecasts were superior to deterministic ones over 4-to-6-day lead-times, even if the probabilistic forecasts were generated using a rougher spatial resolution than the deterministic forecast. They post-processed ensembles with a bias-corection method which was used to increase the spread and improve the uncertainty representation, a technique detailed in section 1.1.1.6.

#### 1.1.1.4 Probabilistic hydrological forecasts

Probabilisitic hydrological forecasting aims to solve the problem of uncertainty assessment plagueing deterministic forecasts. Every element in the hydrological forecasting chain contains uncertainty in various forms, and these must be represented as best as possible to gain insights on the actual uncertainty of future hydrological forecasts. For example, NWPs are simplified models of the atmosphere, and thus they cannot be expected to reflect physics perfectly, leading to structural uncertainty. The NWP model also assimilates observations from satellites, gauging stations, buoys, and other probes, each containing their own measurement uncertainties along with the uncertainty of estimating those values at locations where gauging density is low. Moreover, hydrologic models contain different sources of error and uncertainty since they fail to capture all existing physical processes in a basin, such as soil moisture and channel routing. It is thus essential to make decisions while considering uncertainty and basing

this process on probabilistic principles. It is thus clear that every step in the forecasting chain contributes some level of uncertainty that should ideally be quantified to enable risk-based assessment and rational decision-making (Krzysztofowicz, 1999, 2001).

Probabilistic hydrological forecasting provides a method to assess at least parts of the uncertainty chain. For example, when using a probabilistic weather forecast as input to the hydrological model, the resulting forecast members represent the uncertainty associated to the atmospheric state. Forecasters have also used hydrological multimodel forecasts, where a suite of hydrological models provide independent probabilistic hydrological forecasts and are combined into larger ensembles that also include the hydrological model structural uncertainty (Ajami et al., 2007; Arsenault & Brissette, 2016; Brochero et al., 2011; Devineni et al., 2008; Duan et al., 2007). Others include multi-NWP forecasts, where a hydrological model is fed with probabilistic forecasts from multiple NWP models, resulting in an ensemble forecast that also includes an uncertainty assessment of the weather model structure (Bao et al., 2011; Bogner et al., 2012; Kim et al., 2017; Troin et al., 2021).

Using probabilisitic hydrological forecasts has thus been the method of choice for decision making in the face of uncertain hydrological forecasts (Boucher et al., 2011; Cloke & Pappenberger, 2009; Ramos et al., 2013; Seo et al., 2006). Komma et al. (2007) showed that ensemble forecasts show behaviours similar to those of deterministic forecast error, for example, both have a larger error with increasing lead-time. Therefore, ensemble of forecasts could be used as indicators of potential and/or expected deterministic forecast errors. The hit rate of probabilistic forecasts is also better than deterministic forecasts in issung alarms for flood events (Pappenberger et al., 2005).

Streamflow forecasting is always done in the context of uncertainty. Nester et al. (2012) showed errors related to meteorological forecasts and streamflow simulations. They quantified these errors in various spatio-temporal scales, i.e., for different lead-times and different sizes of catchments. Probabilistic forecasting is not only a way to communicate, but also a way to

address uncertatinties associated to hydrological models and initial conditions (Dion et al., 2021), possibly proposing a choice of hydrological models regarding more skilful forecasts (Devineni et al., 2008). Roy et al. (2017) presented a platform to monitor and forecast streamflow in saparsely-gauged catchments in the context of probabilistic forecasting using multimodel and multi-input methods. They used precipitation from satellite data from multiple sources to force a suite of hydrological models. Dion et al. (2021) utilized several hydrological models in a streamflow forecasting chain to address biased, under-dispersed ensemble forecasts. Devineni et al. (2008).considered two criteria (similarity of past and current preditor state regarding climate variability, besides the performance of streamflow forecasts of each hydrological model) to weighted models in a multimodel setting. They found the multimodel approach was better than the individual model in having fewer false alarms and more reliable forecasts. Li and Sankarasubramanian (2012) made a point about the benefits of using multimodel approaches on monthly streamflow forecasts. They highlighted the effect of multimodel approaches on addressing uncertainty due to measurement errors and structural deficiencies of hydrological models. Verification metrics used to evaluate forecasts affect decision making, and the relative performance of different multimodel structures with varying numbers of models, and different combination strategies (Mendoza et al., 2014). Longer timeseries help to better calibrate multimodel ensembles, especially when models are data-driven (Mendoza et al., 2014).

#### 1.1.1.5 Recent innovations in hydrological forecasting

Hydrological forecasting is a deeply integrated process using a multitude of subprocesses from weather forecasting to streamflow postprocessing. This section presents methods that cannot be cleanly categorized in the abovementionned sections.

Araghinejad et al. (2006) constructed a probabilistic framework with a geostatistical approach to figure out the nonlinearity between streamflow or rainfall variation with large-scale climate indices. It included local regression method using ocean-atmospheric signal and hydrological

condition of the catchment to predict seasonal streamflow. Zhao et al. (2011) compared three types of streamflow forecasting methods (deterministic forecast, deterministic-base probabilistic forecast, and probabilistic forecast) for real-time hydropower reservoir operations. They found that using probabilistic forecasts brought more efficiency for reservoir operations; however, the size of the reservoir storage and the streamflow variability affected the skill of these forecasts.

Madadgar and Moradkhani (2013) compared a Bayesian framework with ESP and found more reliable results using the Bayesian framework. Their framework included copula functions to estimate hydrological drought, i.e., defining correlations between spring runoff and either previous winter runoff, or previous fall runoff.

Finally, Prudhomme et al. (2017) presented an operational setting for streamflow forecasting in multi-temporal scales. They combined three methods, including two basin-scale methods (ensemble streamflow forecasting using historical meteorological data; extracting important, frequent hydrological event on basin-scale) and one regional-scale method (modelling streamflow and groundwater level) to assimilate hydrological conditions and maintain a constant skill. They used different sources of data, including daily rainfall and evapotranspiration observations, daily streamflow, groundwater level observation data from all over the region and ensemble rainfall forecasts.

#### 1.1.1.6 Statistical pre-/post-processing in hydrological forecasting

Using pre- and post-processing methods are essential for addressing errors in model structure and the hydrological model's initial states. Typically pre-processing refers to the act of correcting the characteristics of weather forecasts prior to hydrological forecasting, whereas post-processing refers to the the application of algorithms to correct the statistics of the hydrological forecasts themselves (Lucatero et al., 2018; Troin et al., 2021; Wu et al., 2009). Pre- and post-processing improves forecasts by comparing past forecasts (also known as

hindcasts) to the observations in the past and evaluating the forecast error. This error can be used to condition the operational forecast and improve (i.e., reduce) systematic and conditional biases. It is typically recommended to process the raw forecast data (metorological and streamflow data) (Li et al., 2017) as this can help improve forecast skill. Post-processing methods also were used to choose models in a multimodel setting and constructing ensembles of forecasts while respecting spatio-temporal structures and inter-variable dependencies (Li et al., 2017). Non-stationarity of data could also be addressed using post-processing methods (Ceola et al., 2014). Statistical post-processing faces difficulty while dealing with forecasting extremes; it is a broad issue that needs more investigations (Friederichs & Hense, 2007). (Bellier et al., 2017) also showed that it is possible to implement pre- and post-processing simultaneously, correcting biases in the weather forecasts and resulting hydrological forecasts. Crochemore et al. (2017) bias-corrected precipitation forecasts used for issuing ensemble streamflow forecasts, in order to achieve higher forecasting skill, and Hashino et al. (2007) used a bias-correction method as post-processing for streamflow volume forecasting in monthly and seasonal scales. The scientific literature is rife with examples of pre-processing of inputs used to force hydrological models (Kang et al., 2010) and post-processing of streamflow forecast ensembles (Bellier et al., 2017; Bellier et al., 2018; Verkade et al., 2013; Zalachori et al., 2012).

However, forecast pre- and post-processing requires having access to a large historical database to train the bias-correction algorithms, and can provide forecasts that are not physically possible in some cases (Hamill, 2018; Vannitsem et al., 2020). Furthermore, the application of processing methods can improve forecasts for the wrong reasons, in that the physical forecasting process is not improved and can generate poor results that are masked by the processing step. Therefore, there is also a case to be made in which processing schemes should not be implemented when attempting to improve the core of streamflow forecasting systems as to not mask any deficiencies (Arsenault & Côté, 2019; Troin et al., 2021).

#### 1.1.1.7 Sub-Seasonal to Seasonal (S2S) forecasting

As mentioned previously, different forecast lead-times rely on different sources of inputs (i.e., NWPs for short term forecasts and ESP for long-term forecasts). However, the definition of short, medium and long-term forecasts is blurry and depends on the user and application. Short-term can be seen as below 2 weeks and long-term as above one or two months (Harrigan et al., 2018). Due to the disappearance of the persistence of initial conditions in the atmosphere, there are serious doubts about the possibility of NWP's skill improvement in lead-times beyond two-weeks, even when considering advancement in computing power that led to an enhancement of model prediction skill up to these time frames (Shrestha et al., 2013). This leaves a gap between the short and long-term forecasts that are both too far in the future for the NWPs to have any significant skill, and too close to the present for the initial conditions not to have any impacts such as for ESPs (McInerney et al., 2022). This is known as the weather-climate gap (Bennett et al., 2014; Doblas-Reyes et al., 2013; Mariotti et al., 2019; Vitart & Robertson, 2018). There are studies about improving the quality of streamflow forecasts at these durations, but this necessitates more complicated steps to generate seamless streamflow forecasts (i.e., from short to medium to long-term forecasts) (Hudson et al., 2017; (Hudson et al., 2017; Woldemeskel et al., 2018). A forecasting method designed specifically for these timelines has recently emerged. This approach uses coupled atmosphere-ocean-land general circulation models (CGCMs), and in some cases weather forecasting models, to generate called 'subseasonal-to-seasonal' (S2S) weather forecasts (Yuan et al., 2015) which can be fed to hydrological models to generate S2S streamflow forecasts (Arnal et al., 2018). The primary goal of S2S was to improve the forecast skill in terms of a couple of weeks to better deal with upcoming extremes. This can be done by running a NWP for a longer time period than usual (e.g., multiple weeks) and then driving hydrological models with those, and then exploring the statistics of the hydrological forecasts. It is expected that these methods might provide information on upcoming weather patterns and systems but that the timing of events could be off up to a couple of days for long lead-times. S2S is thus a specific forecasting technique that aims to improve upon ESP for medium-range forecasts. In addition, recent studies point out

incorporating climate information and improvements in the estimation of initial hydrologic conditions as the two main paths to improve seasonal predictability (Mendoza et al., 2017), for which the S2S project can help achieve. For more detail on the S2S project, readers are referred to Vitart and Robertson (2018) and Quedi and Fan (2020).

#### 1.1.1.8 Teleconnections and relations between climate indices and streamflow regime

Climate indices are calculated values used to describe the state of the climate and changes in climate systems. Calculating climate indices requires long-term data time series data, which can come from various sources such as sea surface temperature, precipitation, air pressure, and air temperature. Each climate index corresponds to specific data, region, and period of time, and there are many well-defined climate indices in the literature. Because climate changes much more slowly than weather, tracking climate indices can provide useful information about the state of a region's climate, including trends and potential changes in temperature and/or rainfall (Hurrell et al., 2008; Pappenberger et al., 2015; Wang et al., 2013; Zhang et al., 1997).

Climate indices showed good performance on predictibility of seasonal streamflow forecasts in certain regions which have dominant climate mode(s) and strong teleconnections (Mendoza et al., 2017; Najafi et al., 2012). For example, MacLachlan et al. (2015) reported improvements of seasonal forecast system over the Tropics and West Pacific because of the El Nino–Southern Oscillation (ENSO), and extratropics because of the Arctic Oscillation (AO) and North Atlantic Oscillation (NAO).

Hudson et al. (2017) compared the operational meteorological forecasting system of Australia (POAMA) with a model called ACCESS-S1 (the Australian Community Climate and Earth-System Simulator-Seasonal prediction system version 1). ACCESS-S1 was based on the UK office Glo-Sea5-GC2 (Global Seasonal forecast system version 5 using the Global Coupled model configuration 2; MacLachlan et al. (2015)). ACCESS-S1 had higher resolution and a more recent physics parameterization. They compared both on multi-week to seasonal time

scales, using 23 years of hindcast datasets. ACCESS-S1 reduced spatial biases of the average state of the climate, both globally and regionally. It was capable of capturing important large-scale climate indices over Autralia better than POAMA. It was better in forecasting precipitation, maximum and minimum temperatures on multi-week time scales, though in the seasonal scale results were not very different.

In other words, new methods for improving meteorological forecasts showed improvements on seasonal time scales, but the effectiveness of these methods depends on the variable of interest and the spatio-temporal features (such as the location of the basin and the lead-time). Sohrabi et al. (2021) provided insights for enhancing long-term spring flood forecasting, specifically by examining volumetric bias and peak flows. They utilized a conditioned weather generator based on large-scale climate indices and constructed a linear model to characterize temperature and precipitation anomalies in order to perturb the weather generator at the watershed scale. The study found a strong correlation between large climate indices and temperatures, but a weaker correlation for precipitation. Zhao and Brissette (2022) investigated the effects of three important large-scale climate indices - ENSO, Atlantic Multi-decadal Oscillation (AMO), and the Pacific Decadal Oscillation (PDO) - on temperature and precipitation over North America. They showed the importance of understanding internal climate variability in understanding internal hydroclimatic variability and used the output of the Canadian Earth System Model large ensemble (CanESM2-LE) to do so. They found that ENSO is linked to annual precipitation and AMO is linked to temperature, both over most of North America. PDO showed fewer impacts on both variables.

#### 1.1.1.9 Transitions between short-term and long-term hydrological forecasts

An issue arising when combining both short-term and long-term forecasts is that of the transition. Indeed, an ESP forecast will provide data for every time step between the forecast issue date and the end of the forecast period. If a user also wants to improve the initial portion of the forecast (e.g., the first few days), then a method must be devised to “overwrite” the first

days with those of the NWP. However, mismatches between both products can arise due to different numbers of members and because of differing objectives. For example, short-term forecasts need the timing of impactful events to be correct, whereas for longer lead-times the probabilities of occurrence of events might be the important objective. Combining these multi-timestep according to their objectives requires processing to generate useful forecasts.

Some methods have been devised to provide seamless forecasts over all time steps while incorporating data from all sources. McInerney et al. (2022) proposed a post-processing method to have seamless daily streamflow forecasts without losing performance until the monthly scale. They compared their post-processing method with a non-seamless monthly one. The new post-processing method formulated a residual error model on the daily scale while the old one did it on a monthly scale. Results were comparable, and the new method provided similar forecasts but that were sharper and more reliable than the older one. Therefore, they believed using a method which gives seamless daily streamflow forecast is better than a non-seamless one. McInerney et al. (2020) proposed a multi temporal hydrological residual error (MuTHRE) to improve streamflow forecasts and provide seamless forecasts. They considered three types of errors, seasonality, dynamic biases (to consider the fact that hydrological models are not able to capture changes in hydrological process over very long periods like interannual climate variability), and non-Gaussian. Considering these three errors resulted in improvements in forecasts in three temporal scales, i.e., daily (due to non-Gaussian), monthly (due to seasonality), and yearly (due to dynamic biases). Monhart et al. (2019) improved subseasonal (up-to one month) streamflow forecasts using a mix of NWP and ESP. Arnal et al. (2018) compared two seasonal forecast methods on lead-times longer than one month. They forced hydrological model once with histolrical meteorological observations, and once with the European flood awareness system seasonal streamflow forecasts. Results were comparable, and none of the proposed methods was objectively better than the other one.

Given the above approaches to tackle seamless forecasting, this research project focuses on probabilistic forecasts to explain and convey uncertainties and uses both NWP weather

predictions for short-term forecasts and ESP for long-term forecasts. These forecasting approaches can be used in either of the forecasting frameworks to transform weather forecasts into streamflow forecasts presented in the next section.

## 1.2 Streamflow forecasting methods

Streamflow forecasts requires transforming weather forecasts into streamflow forecasts by simulating the impacts of the weather on the watershed of interest considering the hydrologic cycle and the hydrologic state of the watershed. This section presents two families of methods used to perform this process: hydrological models and artificial neural networks (ANNs).

### 1.2.1 Streamflow forecasting using hydrological models

The most common way of performing hydrological forecasting is using a hydrological model. To do so, the hydrological forecaster runs an already calibrated hydrological model over a period of time preceding the forecast issue date, using observed meteorological data as inputs to the model. This allows warming up the model to a hydrological state that is representative of the actual watershed conditions. This aims to provide a solid foundation for the forecast, ensuring that the forecast is not biased due to an incorrect representation of the watershed state, including soil moisture, snowpack water equivalent and groundwater saturation depth (Kim et al., 2018; Mai et al., 2020). The model is then fed the weather forecast data (from an NWP or from ESPs) one at a time from the same initial conditions such that the only variation between the ensemble hydrological forecast members stems from differences in the weather forecast members. The resulting hydrological forecast can then be used to assess probabilities of attaining a certain streamflow threshold, or to estimate the likelihood of a drought over the course of the forecast period. Forecast skill can be evaluated either in an operational context (with initial state errors that are more representative of actual forecasting conditions, or using pseudo-observations as observed streamflows, which are generated using simulated streamflow instead of using observed streamflow (Harrigan et al., 2018). This forecasting

method removes the need for data assimilation (DA) since the model states are always exactly aligned with the pseudo-observed (i.e., simulated) streamflow. Alternatively, performing operational forecasting using real observed streamflow as the target provides more realism and is more in-line with actual forecasting conditions, at the expense of a more complex error structure since many uncertainty sources combine to alter the forecast performance. Thus, the streamflow forecaster needs to deal with uncertainty originating from initial states to minimize errors when forecasting in an operational context (Sun et al., 2018), as seen in section 1.2.1.1.

The hydrological model-based forecasting approach is the most commonly used one due to the fact that hydrological models are typically easy to understand and represent the physics (or at least attempt to represent the processes in a simplified manner) such that the model responds to various weather inputs during the forecasting process. Hydrological models range in complexity from very simple lumped and conceptual models that provide an estimate of streamflow at the outlet only using simple conceptual process representation (Anctil et al., 2003), to complex physically-based and distributed models that attempt to reproduce the entire hydrological cycle as accurately as possible (Chen et al., 2016). In all cases, hydrological models are imperfect representations of the watershed and of the hydrologic cycle, meaning that they also provide a fair amount of uncertainty and error in the process. Therefore, a hydrological model that is used to perform forecasting might still deviate from the observed streamflow at the time of forecast, even following a spin-up period, which would lead to erroneous and biased forecasts due to the errors in the initial conditions. This drawback from hydrological models can be improved to a certain extent using methods to correct the initial states. These methods are known as data assimilation (DA) methods and are presented in the following section.

#### 1.2.1.1 Data assimilation

DA aims to find the best possible initial state of the hydrological model (e.g., snowpack water equivalent, soil saturation, groundwater levels, etc.) such that the hydrological model

represents the watershed's hydrological conditions as accurately as possible (Wood et al., 2016). This is done by comparing the simulations from the hydrological model as well as the observed streamflows, and, while also considering their respective uncertainty bounds, modifying the model states such that they reduce the error between the modelled and observed flows. This is an important step especially for short-term forecasting because errors in initial conditions strongly affect short-term forecast skill. As an example, we can imagine a scenario where the observed streamflow for a given day is 100 $m^3/s$, whereas the model simulates (i.e., expects) 50 $m^3/s$. Forecasts would therefore be performed with a hydrological model that is underestimating the amount of water in the watershed from the start, leading to streamflow forecasts that are too dry compared to what could reasonably be expected. DA aims to reduce this error and provide the most realistic hydrological states. It can also be a method to assess the uncertainty related to the initial states as most DA methods are probabilistic in nature and provide a distribution of likely initial states that can be used in a probabilistic forecasting framework (Abbaszadeh et al., 2018; DeChant & Moradkhani, 2014; Liu et al., 2012). This in turn leads to more accurate as well as more reliable forecasts. DA updates the initial states sequentially using streamflow observations as targets and progressively bringing the model states closer and closer to the optimal values at each time step. DA methods range from simple to advanced methods (Liu et al., 2012; Troin et al., 2021), which will be highlighted in more details in this section due to the importance it plays in this research.

DA methods are categorized as simple to complex according to their approach towards the filtering problem. The filtering problem includes the state space mapping problem, in which a forecaster tries to decrease errors between the model's image of the hydrological system and the observed image of said system. Errors between the model and observations are caused by input data errors, model structural error and observation error of streamflows. This filtering problem was alleviated by different DA methods (Liu et al., 2012; Madsen, 2003; Sun et al., 2016; Troin et al., 2021). The Kalman Filter is among the most widley used and robust methods (Madsen, 2003; Sun et al., 2016; Troin et al., 2021). The Kalman Filter has been the basis of multiple derivative methods such as the Extended Kalman Filter (EKF; Muluye (2011)), the

Ensemble Kalman Filter (EnKF; Rakovec et al. (2012)), and the Recrusive Ensemble Kalman Filter (REnKF; McMillan et al. (2013)). In Kalman Filter approaches, probabilities of possible states are updated at each time step based on the prior knowledge of the hydrologic state (from the hydrological model) and the measurement (from the hydrometric station). This has the advantage of being robust to outliers and sporadic errors in the measurements or simulations as the next time steps will automatically reconverge to the observations after a certain period. This is a strength for operational settings as well as for research applications where initial states must be generated for long periods of time in a historical time series.

Another DA method that is gaining in popularity is the Particle Filter (PF), which is based on the weighting of "particles" that represent hydrological model states (Weerts & El Serafy, 2006). Each particle thus represents a set of states that perfectly describes the state of the hydrological model. By evaluating the streamflow simulated by each of these "particles", a weight can be given as a function of the error between the observation and the modelled stremflow. Then, particles that are more accurate (i.e., display less error compared to the observation) are weighted more heavily than those that provide poorer simulation accuracy. The worst particles are discarded, and new particles are inserted according to the weighted distribution of the "good particles". These states are then updated by running the hydrological model for another time step using each of these particles as a starting point. The process is then repeated for this time step. The idea is that in time, the better particles will filter to the top and be heavily weighted and represent the initial states distribution very well, while the poorer states are systematically removed and replaced with better ones, leading to a high-quality distribution of initial states. However, this also comes at the cost of robustness. Indeed, particle divergence may happen in some cases where the particles are all of poor quality and cannot be replaced by good quality particles. The opposite situation is also likely to occur in longer simulations, where the entire distribution collapses into a single point, where all particles are identical, leading to numerical instability and the failure of the method to proceed.

Kalman Filter methods have typically been used more than PF as they are more robust and tend to provide excellent results (Mazzoleni et al., 2018). In addition, EnKF have the ability to treat uncertainties explicitly because of their ensemble nature which considers equal weighting for each member (Abaza et al., 2017; Khaki et al., 2017; Moradkhani et al., 2018; Thiboult et al., 2016). Thus, the EnKF is widley used as the DA method for probabilistic hydrological forecasting (Samuel et al., 2019).

Despite the popularity of EnKF methods, they are not strightforward to implement. Some of the challenges hydrologists face include: (1) error determination (Sun et al., 2016), (2) joint state and parameter estimation (Hendricks Franssen & Kinzelbach, 2008; Moradkhani et al., 2005), (3) timing errors (Clark et al., 2008), (4) multi-observational and (5) large-scale applications (Sun et al., 2016; Troin et al., 2021). Calculating hyper-parameters of the EnKF method and selecting which of the hydrological model's internal states to update are other challenges that must be resolved (Moradkhani et al., 2005), because it has been shown that letting the EnKF update all model states at all times is a suboptimal approach (Thiboult & Anctil, 2015). Furthermore, the selection of hyperparameters affects the accuracy and reliability of streamflow forecasts (Bergeron et al., 2021). There is evidence in the literature that EnKF-based DA method is more effective for shorter lead-times (Reichle, 2008; Thiboult et al., 2020; Vergara et al., 2014) since it can update state variables to improve initial conditions, but this can come at the expense of the mass balance of water in the model, leading to poorer long-term forecasts in the process (e.g., removing water from the soil layers for short-term improvements but then underestimating flood volumes at a later time due to this missing water) (Mai et al., 2020). Considering all these challenges, EnKF-based methods still improve forecasts more reliably than PF and others, such as emprical ones (Jiménez et al., 2019) and sequential ensemble-based DA methods (Piazzi et al., 2021).

### 1.2.2 Streamflow forecasting using data-driven methods

Data-driven methods, including regression-based, articifial neural networks or other statistical methods have been used for a long time to predict streamflow on various lead-times (Solomatine et al., 2009; Z. Zhang et al., 2018). However, they often share a similar problem based on the fact that they do not represent the initial state of the catchment accurately and are therefore dependent on streamflow observations of the previous timesteps to condition the forecast. For example, autoregressive models use observed streamflow of previous timesteps in order to forecast the next, but this does not allow including information from weather forecasts or from hydrological processes that are much slower than a few days. For example, a forecast in the spring months will require knowing how much snow is present on the catchment in order to condition the forecasts on the snowmelt if temperatures rise above the melting point. Even classical articifical neural networks (ANNs) were unable to provide much in terms of long-term forecasts for this reason. While they could forecast flows using observed weather and recently observed streamflow and other hydrological variables, they lacked the memory of long-term processes that is required for many hydrological processes, dooming these methods to fail for forecasts of any significant length.

Then, a new type of neural network called Recurrent Neural Networks (RNN) were developed, that were able to hold variables in memory for long periods of time, therefore imitating the hydrologic states of hydrological models (Troin et al., 2021). One such type of RNN is LSTM model (Hochreiter & Schmidhuber, 1997), that can be trained on long sequences of historical data and learn patterns such as snow accumulation and melt, baseflow from saturated groundwater stores and so forth (Shen & Lawson, 2021; J. Zhang et al., 2018). LSTM models have memory cells in their architecture. Memory cells have self-recurrent connections with three gates to manage weights and biases on inputs (input gate), outputs (output gate) and to stop tracking a variable in time (forget gate). In particular, the forget gate enables LSTM models to understand and reflect better on long dependencies that exist in hydrological

processes. LSTMs thus manage to overcome the challenges of short-term memory and vanishing gradients during training which exists in traditional ANNs (Xu et al., 2020).

Furthermore, as opposed to hydrological models, these LSTMs learn the physical processes from the data directly and do not need any information regarding the hydrological processes. Instead, the LSTMs will learn from the input time series and seek patterns that it can represent using a system of weights and biases for nodes that are computed at each time step. While these methods have the ability to be very flexible and learn complex structures from the data, they need large amounts of high-quality data to be implemented (Damavandi et al., 2019). It has been recently showed by multiple authors that using LSTM networks provides realistic simulations of streamflow, oftentimes of higher quality than that of traditional hydrological models (Arsenault et al., 2023; Kratzert et al., 2018; Kratzert et al., 2019). LSTM models have recently been applied in various sub-fields of hydrology (Arsenault et al., 2023; Hu et al., 2018; Hunt et al., 2022; Kratzert et al., 2018; J. Zhang et al., 2018). Potential benefits of using LSTM in hydrology were demonstrated by many researchers in the past few years. For example, J. Zhang et al. (2018) showed the advantage of using LSTM over other RNNs without cell memory for simulating water levels. LSTMs showed better performance than the hydrological model “Sacramento Soil Moisture Accounting Model” (SAC-SMA) in long-term streamflow simulation over more than 200 basins (Kratzert et al., 2018). It was also showed that LSTMs could provide skill in estimating streamflow at ungauged sites (Arsenault et al., 2023); Kratzert et al. (2018); (Kratzert et al., 2019; Le et al., 2019). They tested LSTMs over other regions than they were trained on and showed that they displayed ‘spatial-durability of forecast skill’. Comparing LSTM with Global Flood Awareness system (GloFAS) was done by Hunt et al. (2022) over different climatology regions. They showed better forecast skill for LSTMs up until 5 days lead-time. Also in this context, Hu et al. (2018) showed the better performance of LSTMs over hydrological models, both lumped and process-based, were because of the existence of the forget gate in the LSTM architecture. However, other studies highlighted the drawbacks of LSTMs, such as predicting streamflow during extreme events, which is a challenging undertaking for LSTMs if they were not exposed to such events during their

training. Indeed, LSTMs are excellent when applied to problems they are trained on but are sensitive in extrapolation problems. They therefore cannot be used to replace hydrological models where complex situations that are difficult or impossible to predict by data alone, such as in complex groundwater-river interactions and when human interaction alters the data, such as in dam operation (Khoshkalam et al., 2023; Kratzert et al., 2018; Lees et al., 2021).

## 1.3 Forecast verification

The evaluation of forecasts is done by using different types of metrics and methods. First and foremost, it is important to note that forecasts are typically evaluated in hindcasting mode, whereby the forecasts are evaluated after the forecast period has passed (Oreskes et al., 1994). This allows evaluating the accuracy of the forecast compared to the actual observed streamflows for the forecasted period. In research applications, hindcasts are performed on the historical dataset with archived weather datasets, allowing immediate feedback by comparing the forecasts to the observations on long records at once (Troin et al., 2021). This allows evaluating forecast skill over long periods and thus evaluating forecasting methods by comparing their skill on the same historical periods. Forecast verification is done through a series of metrics, each evaluating a specific aspect of the forecast skill. They are used to communicate and evaluate the forecast performance in a quantitive way (Thielen-del Pozo & Bruen, 2019). Tracking the quality of forecasts over time, evaluating different sources of uncertainties, comparing the quality of different forecast methods, improving the decision-making process and enhacing forecast skill of specific types of events are other important reasons for using forecast verification methods (Alfieri et al., 2014; Demargne et al., 2009).

Verification methods are used to address two main concerns users typically have about their forecasting systems. First, they should quantify the goodness-of-fit between the forecasted flows and the actual observations, known as the forecast accuracy. Second, they should quantify the reliability of forecasts over a long period, i.e., how often is the forecast generally correct and how often does it perform poorly (He et al., 2016). Metrics to do so are numerous,

but one of the most often used is the Mean Absolute Error (MAE) for deterministic forecasts and its probabilistic counterpart, the Continuous Ranked Probability Score (CRPS) (Matheson & Winkler, 1976). The MAE represents the absolute difference between the estimation (or forecast) and the reference/target (i.e., observation), as described in equation 1.1.

$$MAE_t = \frac{\sum_{m=1}^{N} \left| F_{m_t} - O_t \right|}{N} \tag{1.1}$$

where t is the time index of forecast streamflow F, O is the observed streamflow, m represents the member from the ensemble of size N.

The CRPS is a probabilistic generalization of the Mean Absolute Error (Gneiting et al., 2007) and represents the quadratic error between the cumulative distribution function (CDF) of the ensemble forecasts and the unique observation (Alfieri et al., 2014). The CRPS is defined in equation 1.2.

$$CRPS = \frac{1}{T} \sum_{t=1}^{T} \int_{-\infty}^{+\infty} [F(x_t) - F(y_t)]^2 dx \tag{1.2}$$

where T is the total amount of time steps. The CRPS ranges from 0 to + ∞, zero being a perfect forecast. For streamflow, the CRPS unit is m3/s.

A CRPS value can be computed for each lead-time and forecast issue date, therefore it is often represented as the mean CRPS of all forecast issue dates. Typically, CRPS will be unique to each lead-time, meaning that for a 10-day forecast, 10 CRPS values will be provided (each representing the mean of the CRPS values for all forecast issue dates at that lead-time) (Hersbach, 2000). For a probabilistic distribution, the best CRPS is zero (all members have the exact same value as the observation, thus perfect accuracy and no uncertainty) and units are the same as the modeled variable ($m^3/s$ for hydrological forecasts). The CRPS contains

information on the accuracy and also the reliability of the forecasts (Hersbach, 2000). Therefore, probabilistic forecasts with lower CRPS values are more reliable on average than those with higher CRPS values. Other metrics and tools, such as the Brier Score and Rank Histograms, can also be used to help evaluate forecasts where appropriate (Troin et al., 2021), but the MAE and CRPS are the two main metrics used in this thesis. Others will be introduced, when necessary, in the following chapters.

## 1.4 Value of streamflow forecasts in hydropower management

Given that this thesis largely focuses on a hydropower generating watershed, the value of hydrological forecasts in hydropower management must be considered. Indeed, the value of forecasts (and improvements made to them) will vary in space and time. For example, a hydropower system that has a full reservoir and is currently spilling excess water from the snowmelt period will not benefit from a more accurate short-term hydrological forecast since all incoming water will be spilled anyways. Therefore, the value of hydrological forecasts is dynamic and related to the hydropower system state (Arsenault & Côté, 2019; Caillouet et al., 2022; Harou et al., 2009; Ramos et al., 2016; Roulin, 2007).

Forecast lead-time also affects their value depending on the situation. For events like flood prediction, or for small reservoirs or drier-than-usual reservoirs, short-term forecasts are more valuable (Saavedra Valeriano et al., 2010). For large, stable reservoirs with long-term objectives, long-term forecasts are more impactful (Maurer & Lettenmaier, 2004). For example, Arsenault and Côté (2019) evaluated the value of long-term ESP forecasts to a hydropower system and found that adding a small positive bias to the ensemble forecasts was favorable as it forced the reservoir operation to be more aggressive, which counteracted the tendency of the optimisation model to underestimate spilling risks. Boucher et al. (2012) did similar study, where the gains of using ESP were quantified in terms of fewer unproductive spills, more hydropower generation and lowering flood. They used a stochastic decision-making tool and mentioned the positive effects of post-processing the hydrological forecasts

on electricity production. Alemu et al. (2011) presented a decision support system to provide recommendations for sequences and volumes of releasing water. They simulated the operation of a water release procedure from a reservoir and used ESP to optimize the process. They considered optimizing generated electricity while respecting security issues related to flood prevention, as well as environmental and supply concerns. (Anghileri et al., 2016) proposed a framework to quantitatively present the value of ESP forecasts for hydropower management. They studied streamflow ensemble scenarios to optimize water release on a daily basis using synthetic scenarios.

The value of hydrological forecasts therefore requires access to a water release decision support system to quantify. Only through this type of tool, that can estimate energy production and other variables related to the reservoir management, can the value of forecasts and any proposed improvement be quantified. However, this is a difficult process as it requires making drawdown decisions at each time step using complex optimization models. Therefore, in this research, forecasts are evaluated using metrics as described in section 1.3 and their value is not quantified for a specific system. Instead, results are presented in terms of improvement over a reference method and thus the methods presented herein can be applied in any hydropower system.

# CHAPTER 2

# OBJECTIVES AND THESIS ORGANIZATION

## 2.1 Research objectives

The literature review highlighted that forecasting flows is a non-trivial task that requires that many processes be implemented and coordinated such that each part of the forecasting process contributes as much as possible to reliable and accurate forecasts. These forecasts can be used for multiple lead-times, going from short-term forecasts for decisions requiring immediate decisions (such as flood forecasting and evacuation), up to long-term forecasts needed for strategic planning (such as for hydropower management). The main objective of this research project is improving the quality of hydrological forecasts on multiple lead-times, such that they can be more useful to stakeholders and water resources managers. To attain the main objective, the following secondary objectives were defined:

1) Improving long-term hydrological forecasts using climate indices and other filtering mechanisms to better select members from ESP ensembles.
2) Improving short-term forecasts by developing a large-scale DA implementation for a distributed hydrological model.
3) Improving short-term hydrological forecasts and evaluating the impacts of DA using LSTM models.
4) Quantifying the uncertainty of streamflow forecasts such that the uncertainty can be communicated more efficiently.

## 2.2 Thesis organization

This thesis is based on three scientific papers either published or submitted to internationally recognized and peer-reviewed journals, each targeting one of the secondary objectives and, the

main objective in their aggregate. Each of these papers is presented as a chapter in this thesis, as follows.

Chapter 3 presents the first paper titled *Evaluation of the potential of using subsets of historical climatological data for ensemble streamflow prediction (ESP) forecasting*. It was published in the Journal of Hydrology in April 2021. This study aimed to improve streamflow forecasts on long lead-times using the ESP methodology. To do so, methods to process and select members from ensembles of scenarios were explored. Optimizing the selection of forecast scenarios considering the potential within the histroical climatological data was investigated. A novel way of calculating the potential skill of ESP-based forecasts for long-term (e.g., more than one month) was proposed. The significant contribution of this paper to the literature is the development of a novel method to evaluate the maximum possible skill that can be extracted from an ESP forecast, thus setting bounds on possible expected performance of ESP-based forecasts. It also allowed exploring some of the relationships between climate indices/teleconnections and long-term streamflow forecasts, to better understand and predict streamflow using the natural, low-frequency, and internal variability of the climate system.

Chapter 4 presents the second paper titled *Sensitivity analysis of the hyperparameters of an Ensemble Kalman Filter application on a semi-distributed hydrological model for streamflow forecasting*. It was submitted to the Journal of Hydrology in May 2022. The aim of this study was to improve forecasts in shorter lead-times by implementing an EnKF DA strategy on a semi-distributed hydrological model. It allowed improving hydrological forecasts on multiple time steps up to 9 days by implementing a hyperparameter selection algorithm developed in this study. A sensitivity analysis was used to find the best state variables to update for each season along with the specific uncertainty assessments to implement. This paper's significant contribution is the implementation of a spatially distributed DA method and sensitivity-based state variable selection method.

Finally, Chapter 5 of this thesis presents the third paper titled *Comparing a long short-term memory (LSTM) neural network with a physically-based hydrological model for streamflow forecasting over a Canadian catchment*, which was submitted to the Journal of Hydrology in December 2022. In this paper, short-term forecasts are once again targeted but they are performed using an LSTM neural network model. The LSTM model is compared to a traditional semi-distributed hydrological model for forecasting. The main contribution of this paper is the fact that the initial states for forecasting are fed to the LSTM as inputs in the form of the observed streamflow at the time of forecast. This model therefore does not require DA as it uses the observed streamflow at the time of forecast to adjust the forecast amplitude. The LSTM model is compared to the hydrological model, both with and without DA implementations, and the LSTM is shown to outperform the hydrological model in most lead-times.

# CHAPTER 3

# EVALUATION OF THE POTENTIAL OF USING SUBSETS OF HISTORICAL CLIMATOLOGICAL DATA FOR ENSEMBLE STREAMFLOW PREDICTION (ESP) FORECASTING

Behmard Sabzipour [a], Richard Arsenault [a] and François Brissette [a]

[a] Hydrology, Climate and Climate Change Laboratory, École de technologie supérieure, Montréal , Canada.

Paper published in Journal of Hydrology, April 2021

## Abstract

Streamflow forecasting is a crucial task for hydropower operations, flood forecasting and water resource use optimization. Ensemble Streamflow Prediction (ESP) is commonly used in the case of long-term lead-times (a few months or seasons ahead). ESP covers the use of historical meteorological scenarios in driving a hydrological model to generate an ensemble of possible future streamflows. Many studies have evaluated methods for selecting optimal subsets of scenarios to improve forecasting skill, and indeed, this is still an ongoing area of research. In this study, we propose a procedure that calculates the maximum potential skill of a classic ESP forecast. A genetic algorithm (GA) is used to determine the best possible set of climatological scenarios given any ensemble size. Along with providing a direct estimate of the ESP forecasting potential in hindcast experiments, the method can be used as a reference for comparing other methods to ESP. The procedure is also used to compare classical ESP, a well-established forecasting method, with two new methods, namely, the Analogue method and the Contingency Table (CT) approach. A discriminant analysis is finally implemented to attempt to identify key features of ESP members that performed well as compared to their counterparts using historic climatology and climate indices. It is shown while there exists a potential for improvement, a lot of research must still be realized to exploit this potential. The procedure

was tested over two basins in Canada. In general, results showed that for any forecast date, decreasing the ensemble size led to a higher potential for better forecasting skills. However, the method does not yet allow identifying the subset of the entire climatology to be used to maximize the ESP forecast performance.

**Keywords**:
Ensemble Streamflow Prediction; Genetic Algorithm; Potential skill; CRPS

## 3.1 Introduction

Hydrological forecasting is important for enhancing the efficiency of water infrastructure management and agriculture, as well as for expanding the hydropower sector (Clark & Hay, 2004; Gutiérrez & Dracup, 2001), and decreasing losses due to hydrological phenomena, such as droughts and floods (WMO, 2009). Typically, hydrological forecasts are generated by driving a hydrological model with one or multiple weather forecasts for the upcoming days or weeks (for short- and medium-term forecasting) to months, seasons, and even years, for long-term forecasting using historical weather scenarios. If a single meteorological scenario is used to drive the hydrological model, the outcome is a deterministic forecast streamflow event. However, this does not allow assessing the uncertainty brought upon the decision-making process by the deterministic forecast. Therefore, driving the hydrological model with multiple meteorological scenarios is a preferred alternative, and returns a probabilistic ensemble of possible future streamflow scenarios, even though in some instances, probabilistic forecasts are more difficult to interpret. It has been almost two decades since (Krzysztofowicz, 2001) mentioned the benefits of probabilistic forecasts over deterministic forecasts. For instance, probabilistic forecasts are able to express uncertainty, make risk-based decisions, and better communicate this information to end-users. In addition, since hydrological forecasts inherently deal with unknown future events, it makes more sense to convey forecasts in a probabilistic manner rather than in a deterministic one. Another issue relating to hydrological forecasting is the presence of different sources of uncertainty, including hydrological model structure and

parameterization, forcings, and initial conditions (Moradkhani & Sorooshian, 2009). All these sources are accompanied by the nonlinearity and complexity of the atmospheric system, thus, making it imperative to implement probabilistic forecasts (Cloke & Pappenberger, 2009).

### 3.1.1 Options for hydrological forecasting

For short-term streamflow forecasting, numerical weather prediction (NWP) can be used, as weather models are known to be skillful to up to approximately 10 days, depending on the region (Cuo et al., 2011; Zhao et al., 2012). For longer periods, NWP becomes unreliable and ESP method provides a baseline for further hydrological forecasting lead-times (Monhart et al., 2019). The simplest method for use in generating ensemble streamflow involves sampling historical streamflow timeseries directly from the historical record. However, this method disregards the impacts of the initial hydrologic state. To overcome this limitation, ESP using climatology for streamflow forecasting was introduced by Hirsch et al. (1977) and (Day, 1985). In this framework, a hydrological model, initialized with observed meteorological data, is forced with historical climatological data to produce streamflow scenarios for the desired forecast period, with each scenario being referred to as a member, and an ensemble consisting of multiple members (Harrigan et al., 2018). In this study, the ESP method is implemented, and the term “ensemble” is used to describe the multitude of climate scenarios, as well as the resulting possible streamflow scenarios. In ESP, forecasts are bounded by historical weather measurements, and therefore, any extreme weather event that is not already present in the historical database cannot be forecasted adequately. However, they do provide information where none would be available otherwise.

### 3.1.2 ESP strengths and limitations

ESP is widely used in different forecasting centers throughout the world (Buizza et al., 2018), and has been considered as constituting the baseline for future forecasting techniques. (Harrigan et al., 2018) showed the high streamflow forecasting skill of ESP over 314 basins in

the UK, issued for lead-times varying from one to six months, although its performance varies by catchment, and is a function of various factors, such as the amount of base flow and the initialization month.

While ESP is the reference method for seasonal forecasting (Monhart et al., 2019), its skill is still relatively limited for longer-term forecasts (Lucatero et al., 2018). It also contains biases and inconsistencies in its skill (Harrigan et al., 2018; Mendoza et al., 2017). These limitations are mostly attributable to the fact that ESP does not consider any information other than historical climatology, nor does it allow conditioning of the forecast on expected weather in the near future. While ESP assigns the same probability of occurrence to all members, it is clear that both wet and dry scenarios cannot occur simultaneously, and therefore, the forecasting skill could potentially be improved by including the current state of the atmosphere in ESP forecasts. In addition, the ESP method cannot take into account climate variability, either internal or anthropogenically forced (which is a drawback in terms of climate change), nor can it forecast extreme events that were not in the historical climatology.

Other methods have been proposed in the literature to provide more reliability and accuracy than ESP. For instance, Arnal et al. (2018) employed seasonal climate forecasts for seasonal streamflow forecasting rather than using historical meteorological scenarios. They concluded that, on average, in the first month of lead-time, their method was more skilled than ESP. The method's performance varied depending on the region and forecast issue time, and the authors concluded that a better understanding is needed in terms of the link between hydrological and meteorological variables. Moreover, Emerton et al. (2018) presented GloFAS (global-scale operational seasonal hydro-meteorological forecasting system) as a worldwide streamflow forecasting system. This system is most suitable for cases where no other forecasting method is applicable, such as in ungauged basins. However, notwithstanding some promising results for lead-times up to 4 months, it showed an over-prediction tendency and, in some cases, was less skillful than ESP.

### 3.1.3 Methods proposed to improve ESP forecasting skill

It is generally expected that adding information (i.e., either adding members or subsetting higher-quality members) to ESP forecasts should enable improvements over the basic ESP framework, as the probabilistic nature of the method then has more information allowing it to adequately determine the uncertainty of future streamflow. For further improvements, both pre-processing (processing of inputs to the hydrological model; Kang et al. (2010)) and post-processing (direct adjustment of the streamflow ensembles; Zalachori et al. (2012), Verkade et al. (2013)) have been proposed, but they typically rely on all members, including those that could possibly be considered non-realistic depending on the hydrometeorological conditions prevalent on the forecast issue date. Furthermore, some challenges are encountered when trying to rebuild meteorological and hydrological scenarios after processing the forecast distributions. Nonetheless, some headway is being made in rebuilding appropriate scenarios from distributions (Bellier et al., 2017; Bellier et al., 2018), whereas ESP is de facto a scenario-based method.

Finding appropriate transfer functions between expected future climate states (probability of being in a wet or dry period, amount of precipitations, etc.) and ESP member weighting is a challenging task. For example, the natural climate variability induces a great deal of uncertainty in the prediction of future streamflows. One way climate variability can be considered in ESP is by taking into account large-scale climate indices, such as the El Nino–Southern Oscillation (ENSO) and the Pacific Decadal Oscillation (PDO) (Grantz et al., 2005; Hamlet & Lettenmaier, 1999). Studies have been conducted that point to a correlation between some climate indices and streamflow, or with other related factors, such as the Snow Water Equivalent (SWE) and precipitation. For instance, ENSO is one of the factors that influences precipitation patterns over western North America. In the case of Western North America, wetter winters in the south and drier winters in the north occur more frequently during El Niño's warming phase (Dettinger et al., 1998; Trenberth, 1997; Zhang et al., 2010). In addition, McCabe et al. (2004) showed that PDO and the Atlantic Multi decadal Oscillation (AMO)

account for 52% of the total variance of drought frequency in the U.S. Such correlations and connections convincingly lead to the conclusion that climate indices can be considered as uncertainty-reducing sources for streamflow forecasting. Therefore, some studies have proposed methods that take advantage of teleconnections and large-scale climate indices to improve climatological forecasting through ESP.

In their case study, Najafi et al. (2012) considered correlations between climate signals and streamflow as predictive factors. They considered different weights for ESP members, and for each forecast date, they chose different sets of climatological scenarios. Their results were promising for predicting spring runoff. Beckers et al. (2016) considered ENSO to distinguish between different ESP traces. They selected historical scenarios according to the similarities between the ENSO index in the forecast year and the historic year. They then generated more scenarios stochastically by conditioning on ENSO to compensate for data removed in the previous phase. They found a 5-10% skill improvement in two of three case studies. While both of these studies showed improvements in long-range forecasting, neither could evaluate how much skill could have been gained if an ideal method were to be used and how close their methods got to achieving this target, with an ideal method being one that would return the best possible forecast skill given the ESP members on hand. Simpler methods, such as in Yao and Georgakakos (2001), were also considered. The authors introduced the Analogue method, involving selections from historical traces. In this method, traces are ranked according to the Euclidean distance between past and inflows at forecast date. Based on this proximity method, they selected the most similar historical traces as probable realizations of future streamflow. Results for March-June were especially satisfying in terms of forecasting biases. Koutsoyiannis et al. (2008) also used the analogue method in a trial-and-error procedure to determine the best length of the backward-looking predictor period and the number of historical traces used in the ensemble. Similarly to this study, Svensson (2016) proposed river flow forecasting using hydrological persistence and historical analogues, and found that the analogue method performed better for longer lead-times. The analogue method was cited as a possible benchmarking method for further studies.

### 3.1.4 Study novelty and objectives

The current literature includes many studies which address methods to improve ESP, and results are often compared to classic ESP results to determine improvements brought by proposed methods. In this study, 'classic ESP' refers to the ESP method, without any modification, and considering that all historical scenarios have an equiprobable chance of occurring in the forecast period. However, to our knowledge, no study has attempted to determine the theoretical maximum skill that could be achieved through ESP methods, either quantitatively or qualitatively.

This so-called theoretical maximum skill is an equivalent of maximum forecast skill which could be gained by using historical scenarios in ESP. This information could help extract as much forecasting accuracy as possible from ESP forecasts, as well as to identify some member properties that could be useful for differentiating members that will lead to better predictions from the rest. Thus, the aim of this study is to quantify the theoretical maximum skill present in historical meteorological scenarios for long-term streamflow forecasting, given any ensemble size. A secondary objective is to determine how to best identify a subset of available members to produce a skillful ensemble for ESP forecasting.

## 3.2 Data and case studies

This study was performed on two catchments in Canada, one in the province of Quebec (QC) (Matawin River, hereafter referred to as the 'QC catchment') and the other in the province of British Columbia (BC) (Chilko River, hereafter referred to as the 'BC catchment'). Both catchments are snowmelt-dominated and have warm and cold seasons with similar precipitation patterns, as shown in Figure 3.1.

The hydrometeorological data were taken from the CANOPEX hydrological database, which contains hydrometeorological data for 698 catchments across Canada (Arsenault & Brissette, 2016). The observed weather data (daily precipitation, maximum and minimum temperatures) in CANOPEX are sourced from the Natural Resources Canada gridded climate database (Hutchinson et al., 2009). Hydrometric data in CANOPEX comes from Environment and Climate Change Canada's Water Survey Canada (WSC) hydrometric database. Climate data covers the 1950-2010 period, while hydrometric data includes all available observations over the same period. For this study, given the need to generate ensembles based on historical meteorological data, two catchments were selected covering the complete 61-year study period 1950-2010.

Four climate indices were also employed to attempt to estimate future flow and precipitation. These are the Pacific-North American (PNA), Multivariate El Niño/Southern Oscillation Index (MEI), Pacific Decadal Oscillation (PDO), and Southern Oscillation Index (SOI). These data are publicly available from the National Oceanic and Atmospheric Administration's (NOAA) website: https://psl.noaa.gov/data/climateindices/list/).

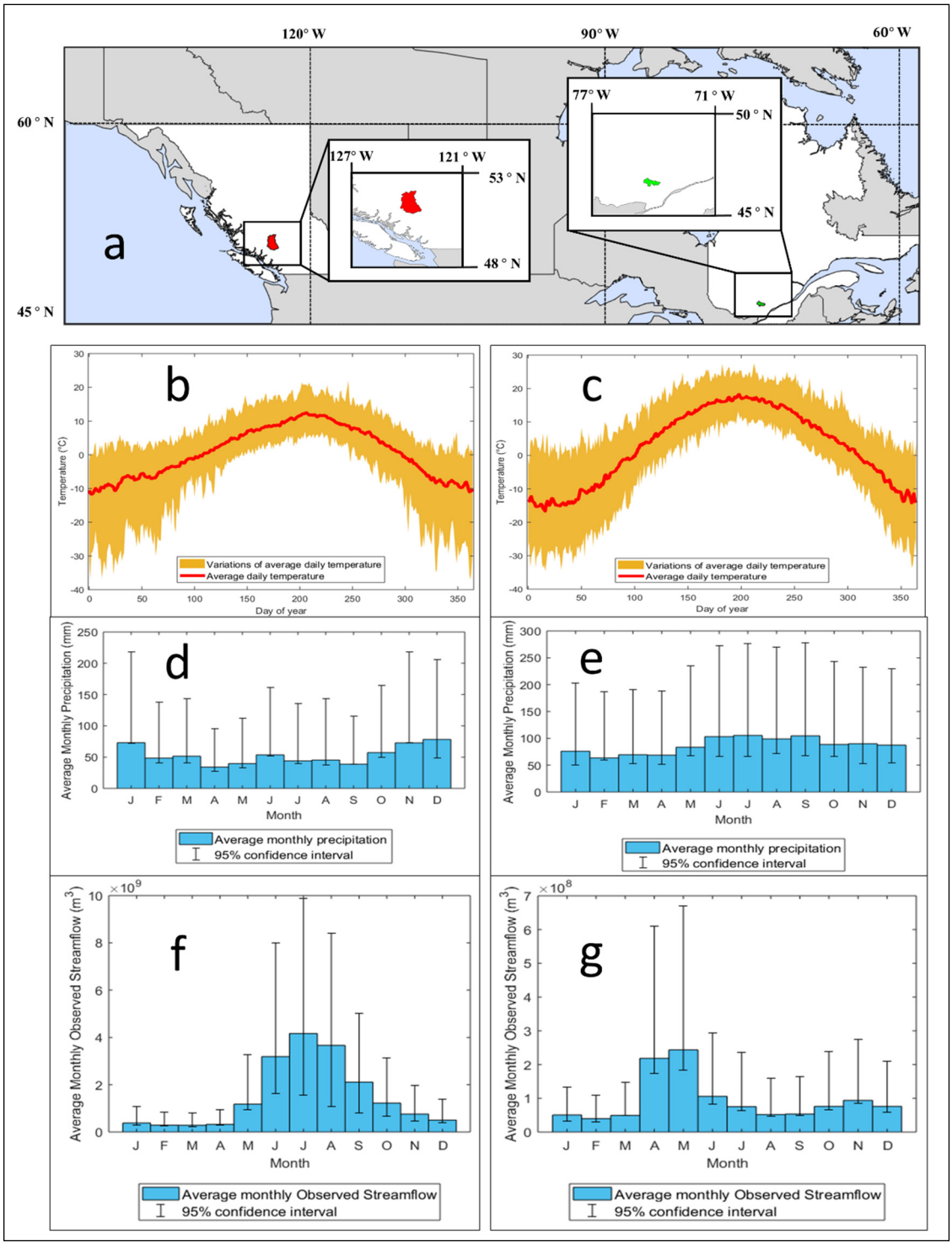


Figure 3.1 Catchments used in this study are shown in panel a). The left-hand panels (b, d, and f) represent the average temperatures, precipitation, and

streamflow in the BC catchment. The right-side panels represent the same variables but for the QC catchment.

## 3.3 Material and methods

### 3.3.1 Hydrological models

For this study, a hydrological model was employed to generate ensemble streamflow forecasts from historical climate data. The hydrological model selected for this purpose was HMETS (Hydrological Model École de Technologie Supérieure) (Martel, 2017), a lumped conceptual model that has been shown to perform well in a wide range of hydroclimatic conditions, including in snowmelt-dominated regions (for instance, it was used by Troin et al. (2015) for North American catchments). HMETS has 21 free parameters, 10 of which are related to the snowpack melting and refreezing processes, 6 to horizontal water routing, 4 to vertical water balance, and 1 to evapotranspiration scaling.

As inputs, HMETS needs daily maximum and minimum temperature data, as well as daily solid and liquid precipitation and potential evapotranspiration, which has been estimated by the Oudin formula (Oudin et al., 2005).

### 3.3.2 Meteorological data processing

Precipitation data was first divided into snow and rain using a linear interpolation based on the mean daily temperature. For days with mean temperature above 3°C, the precipitation was considered as rainfall, while precipitation during days with mean air temperature below -1°C was considered entirely as snow. For days when air temperature averaged between -1°C and 3°C, the fractions of liquid and solid precipitations were linearly interpolated according to that same scale.

### 3.3.3 HMETS calibration

For both catchments, HMETS was calibrated on the entire hydrometeorological time series using the Covariance Matrix Adaptation – Evolution Strategy (CMAES, Hansen et al., 2003). The full available period was used as it allows maximizing the information content in the parameter set (Arsenault et al., 2018). CMAES was given a budget of 15,000 model evaluations to converge to an acceptable solution, and the Nash-Sutcliffe Efficiency (NSE) calibration was 0.60 for the BC catchment, and 0.80, for the QC catchment.

### 3.3.4 Synthetic streamflow

HMETS was first driven by the historical meteorological data to generate a long-term simulated streamflow series. This synthetic streamflow was considered as a proxy for the observed streamflow to remove errors in the hydrological model initial conditions. Many studies have used this method to overcome the need to compensate for model drift (Alfieri et al., 2014; Arsenault & Côté, 2019; Harrigan et al., 2018; Pappenberger et al., 2015). The first year of simulated discharge was removed as the results were unstable due to the hydrological model warm-up.

### 3.3.5 Forecast verification

In the present study, CRPS (Hersbach, 2000) was used to evaluate the forecast skill against the corresponding pseudo-observed streamflow. CRPS is one of the recommended skill scores for forecast verification (Pappenberger et al., 2015), which provides information on the forecast reliability and resolution (Hersbach, 2000). We focus on the accumulated inflows on long horizons, as this is an important criterion when evaluating reservoir drawdown rates and long-term production capacity in hydropower systems. CRPS is a quantitative technique that measures the accuracy of the forecast by calculating the area between the prediction CDF and the CDF of observations (Zamo & Naveau, 2018). It is important to note that the ensemble size

can thus play a role on the CRPS value, with larger ensembles providing lower CRPS values, given the ensembles have the same statistics (Ferro et al., 2008). Lower values of CRPS signify that the ensemble members are closer to the target value (in this case, the accumulated streamflow).

### 3.3.6 Proposed procedure to calculate the potential of using climatological data in ESP

In the ESP method, all members are considered equiprobable. This consideration is simplistic, but in the face of uncertainty, if no other discriminating factor is available, it becomes difficult to provide a means of weighting or selecting members. However, as some previous studies suggest (Bojariu & Gimeno, 2003; Mudelsee et al., 2003), it could be possible to extract information on various scales from autocorrelations in inflow structures, which could then be used to inform on future streamflow characteristics. This would allow weighting some members higher than others, based on a probabilistic view. In this step of the study, the aim is to maximize the benefits of using particular subsets of the ESP forecast members for hydrological forecasting. We seek to obtain the best possible performance, i.e., the smallest CRPS value, for any ensemble size **m**. The target of the hydrological forecasting in this study is to predict the cumulative amount of inflows (CI) at the end of the forecast period. An experiment was designed to evaluate the possible gains in performance based on the available members. In other words, for each possible ensemble size **m**, a series of possible combinations of **m** members without replacement were generated and the CRPS computed. This allowed estimating the shape of the ESP forecast response domain.

It is worth noting that the ensemble size must have at least one member less than the Full-ESP ensemble to allow sampling of possible combinations of ESP subsets for comparison with the Full-ESP basic scenario. The objective here is to verify the possibility of obtaining a better CRPS by using a better selection of members than by using the full ensemble.

To find the best set of **m** members, it was necessary to compare different combinations of members from the full set of **n** members and compare their resulting CRPS. The study was hence performed in hindcast mode to allow retroactive verification of the forecast skill. Finding the best combination of **m** members from an ensemble of size **n** may be considered to be a deterministic binary optimization problem, where each member is attributed a value of 1 if it is selected and 0 otherwise. A black-box optimization tool using the Genetic Algorithm (GA) (Whitley, 1994) was used to find the best combinations of members for all values of **m**. The optimization aims to minimize the CRPS using a binary selection, i.e., each parameter can be set to either 1 (select the given member) or 0 (do not include the member in the ensemble), with one parameter per available member. The only constraint was to force the GA to select a given number of members in order to determine the best combination of members if one were to use 1, 2, 3,… *n* members in the reduced ensemble. The algorithm was run independently for each value of **m**, as opposed to a single, multi-objective optimization, which would have provided the multi-objective Pareto front. The mono-objective approach was chosen in order to better characterize the response landscape and not omit any points that could be dominated from an objectively "better" score in the multi-objective space. The main caveat to this approach is that the results are dependent on the convergence skill of the algorithm in the high-dimensional parameter space (e.g., selecting 25 out of 45 possible members means there are more than $3\text{x}10^{12}$ possible combinations). Therefore, GA can help converge towards the optimum solution, but the actual optimum might be left undiscovered (Wu and Chau, 2006). Nonetheless, this method provides a more targeted approach than a brute-force one. The potential of an **m**-member ESP ensemble forecast would be calculated as follows:

$$\Delta CRPS_m = CRPS_{Full\ ESP} - CRPS_{GA_m} \tag{3.1}$$

where $CRPS_{Full\ ESP}$ denotes a CRPS value corresponding to considering all members in the forecast, and $CRPS_{GA_m}$ denotes the smallest CRPS while considering **m**-members as determined by the GA algorithm. A positive value of $\Delta\text{CRPS}_\text{m}$ ('Delta') indicates that there is

potential for improving the forecast by selecting a subset of ESP members rather than including all the scenarios. For example, in Figure 3.2**,** the skill of the Full-ESP ensemble is compared to that of the best performing combination of **m**-members with **m** ∈ (1, 2, …, **n)** in terms of CRPS for the QC catchment and any lead-time and forecast date for illustrative purposes. The red star indicates the CRPS for the Full-ESP, while the blue circles show the CRPS for different ensemble sizes obtained by the GA method. The horizontal red line indicates the potential to improve the forecast skill for each ensemble size. The difference between the Full-ESP and the GA-derived ensemble with the lowest CRPS is the potential improvement which could be obtained with the available members.

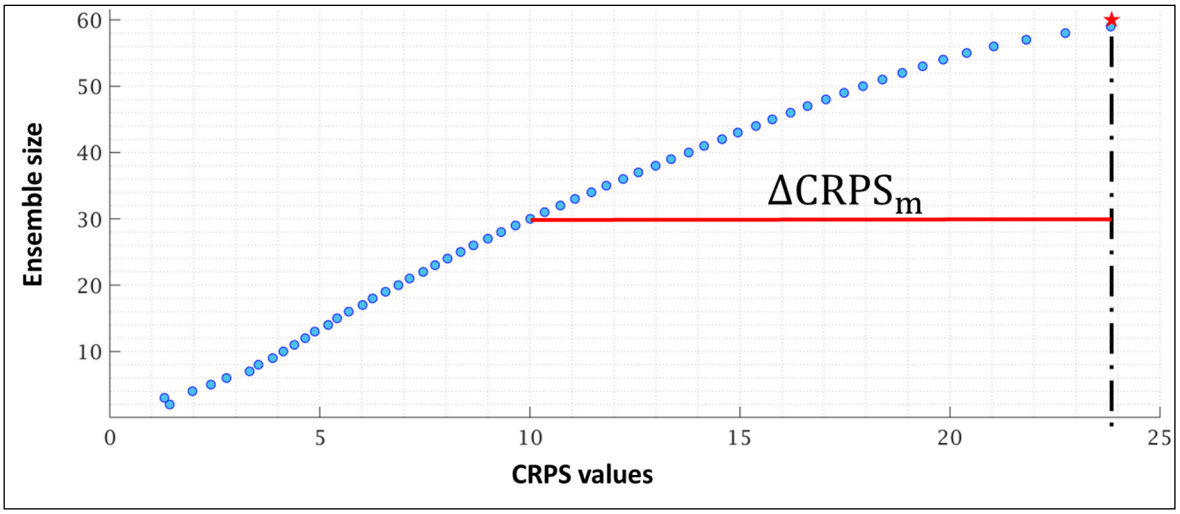


Figure 3.2 The red star is the CRPS value of ESP result for 1st of June 2000, 90 days lead-time, for QC catchment. The blue circles are theoretical optimum CRPS values for different sizes of ensembles. The red horizontal line is ΔCRPSm

Figure 3.2 shows an example of the GA optimization for different numbers of members in the ensembles. There is a clear trend as the the size of the ensemble is decreased that the best possible CRPS values improves as well. The results of the GA approach in Figure 3.2 are somewhat unsurprising, as it is expected that selecting members with full knowledge of the target during optimization should lead to better results. The question then becomes how much of the $\Delta CRPS_m$ value can be extracted by using information available to the modeler at the time of forecast. Essentially, how can this information be used to translate the potential into

real forecasting skill? One possible approach would be to attempt to identify the factors that discriminate desirable members from undesirable ones. Here, we implemented two methods to refine the member selection based on information available to the hydrologist during the forecast.

### 3.3.7 Contingency Table (CT)

In this method, we assumed that the status of the precipitation or discharge tested over a period preceding the initial date of forecast is the most important factor for selecting ESP members (climatological scenarios). The method is based on a historically-driven contingency table (CT). It then goes on to propose considering a certain period of variable length prior to the forecast date, and then using the period's hydrometeorological conditions as predictors for the forecast period, as was done in Hwang and Carbone (2009), Svensson (2016) and Madrigal et al. (2018). The method was implemented as follows:

1- Choose two periods: the first is the period before the forecast date, which will serve as the predictor period; the second is the forecast duration, simply called the forward period.
2- Choose the variable which is going to be forecasted (precipitation, streamflow, or other) at a daily time-step.
3- Calculate a representative value of the chosen variable over the predictor period (for example, average precipitation), which serves as the predictor to the model, for all years except for the year of the forecast date.
4- Discretize the predictor values from the previous step into *p* categories. We classify the data into three categories (p=3):
    a. *Wet*, if the value is in the top tercile,
    b. *Normal*, if the value is located in the second tercile,
    c. *Dry*, if the value is located in the third tercile.
5- For each value in step 3, also compute and categorize the historical observed inflows during the forward period following the predictor period. For example, for a given year,

the predictor could be categorized as *Wet* and the forward period that immediately follows, which corresponds to the forecast duration of the forecast date, could be categorized as *Dry*. These are the targets that will be used to define and populate the different cells of the CT. Note that the target values are computed on the same initial date of forecast, but on all the other available years in the dataset.

6- Build the CT by classifying the data into a 1$^{st}$-order 3-by-3 conditional transition matrix between the predictor period state and the forward period state. This will allow identifying probabilistic transitions from the state in the predictor period to the forward period. For example, what would be the chance of having a *Wet* period after a *Dry* period?

7- Estimate and classify the predictor period value for the forecast date (the period right before the initial forecast date).

8- From the CT, establish the probabilities of having a wet/normal/dry forecast period based on the predictor period state.

After these steps, it is possible to select a subset of members of the Full-ESP ensemble that best represent the probabilities of the future states by random sampling from the appropriate classes. For example, if CT predicts a 60% chance of the forecast being wet, then build a new ensemble that contains 60% of wet scenarios from the available set through random sampling of the wet years and dry years independently.

### 3.3.8 Reassemble ESP with 'Perfect' information about the forecast period

In the last method, we hypothesized that characterizing the predictor and forward periods according to their relative average precipitation (or discharge) would improve the forecast skill over using the entire set due to the resulting improved probability representation. To test this claim, in this section, we introduce a variant we name the CT-Perfect Information ESP (CT-

Perfect). CT-Perfect refers to the idea that the exact state of the forecast period is assumed to be known, so only the members that are from the same category (wet/dry/normal) are used in the forecasting step, without considering the predictor period state. Consequently, this method is an extension of the CT method, but which does not require a probabilistic assessment of the future state, as the method uses otherwise unknowable information for its discrete and perfect classification (i.e., the future period, based on the precipitation variable, is going to be *Wet* with complete certainty). We introduce this method as a benchmark to evaluate the capabilities of the CT method under uncertainty.

### 3.3.9 Analogue method

The analogue method assumes that the best predictor of the forecast period is the predictor period immediately preceding the forecast. In this case, the best ESP member to select as the future realization is the one whose predictor period hydrometeorological properties are most similar to those of the forecast date. This scenario is then selected as a member of the improved and reduced-size ESP. Members can be added to the forecast predictor period in descending order of similarity. The similarity can be measured using a so-called "analogy factor" that consists of two features: 1) hydrometeorological properties (precipitation or discharge), and 2) a statistical aggregation metric of the hydrometeorological variable (i.e., the sum, average or maximum value). The process can be implemented following these steps:

1- Define a predictor period as for step 1 of the CT method.
2- Choose a predictor variable (precipitation, discharge or other). In this study, the discharge was selected for this step.
3- Calculate the accumulated value of the chosen variable (discharge) over the predictor period, for all years serving as predictors to the analogue method.
4- Calculate the absolute difference between the accumulated simulated discharge for all predictor periods (except for the year of the forecast date), and the accumulated simulated discharge for the period before the actual forecast date ($\Delta_i$, with $i$ indicating the corresponding ensemble member).

5- Rank $\Delta_i$ ($i$=1: n) in ascending order.

Choose the first **m** members from the ordered set from step 5 and use the climate data from their forecast periods as inputs to the hydrological model to generate the new ESP for the real-time forecast.

### 3.3.10 Random selection of ESP forecast members

A random sampling from the available members was also investigated as a baseline method. For each value of **m** ∈ (1, 2, …, **n**-1), with **n** being the number of available years to select from, we sampled **m** members randomly to compose the ESP forecast. This process was repeated 1000 times for each value of **m**. The choice was made to select 1000 random samples for each value of **m** as the distribution of the results obtained always converged before that value. Therefore, the sample size was sufficient to explore the response space.

### 3.3.11 Exploration of member characteristics based on forecasting performance

The results obtained through the GA method were analyzed to attempt to determine whether there were any characteristics they shared in common that could help identify the members to select when sampling in an ESP forecasting framework. This analysis is similar to a discriminant analysis, where the members composing the optimal ensemble (hereafter referred to as 'best members') are compared to ensembles of the same size containing the worst members (i.e., the ones that generate an ensemble that provides the worst CRPS), according to a set of hydrometeorological characteristics. For this analysis, the combinations of five members (out of all available members) that returned (1) the best and (2) the worst results in terms of CRPS were identified and investigated to determine how they differed, in hopes of shedding light on the reasons behind their opposite behaviors in forecasting. The best members were selected from the five-member GA optimization solution, while the five worst were taken

from the random sampling that returned the worst CRPS, and that did not include members that were in the best five-member ensemble. The comparison between the two groups was performed based on the following four indices over the predictor periods:

1. The relative difference between mean values of temperature and precipitation of the forward period and predictor periods of each of the best (worst) members.

$$relative\ delta\ for\ \boldsymbol{Mean}\ \alpha_i = \frac{\bar{\alpha}_i - \bar{\alpha}_{forward}}{\bar{\alpha}_{forward}} \tag{3.2}$$

where $\alpha$ denotes the daily temperature or precipitation, either during the forward period (mean value: $\bar{\alpha}_{forward}$) or during the period denoted by the best (or worst) member $i$ (mean value: $\bar{\alpha}_i$).

2. The same as (1), but using the standard deviation instead of the mean of variable α.
3. The correlation between temperature and precipitation data of the forward period and each of the best (worst) members.
4. The relative difference between the average of the climate indices (average and standard deviation of both precipitation and temperature) during the forward period and that of the best (worst) members.

This methodology was developed to attempt to reverse-engineer the characteristics of the members that lead to good and bad forecasts, rather than attempting to determine the explanatory variables *a priori*.

## 3.4 Results

The results in this section are first presented for a single forecast case and are then extended to integrate the notion of reliability of the ESP forecasting methods and to analyze their statistics.

In the first case, forecasts were issued on July 1st, 1990, with a 90-day lead-time. The predictor period was also set to the 90-day window ending on the forecast date of July 1st. Different methods implemented in this study were then analyzed and compared according to CRPS, as can be seen in Figure 3.3 (QC catchment) and Figure 3.4 (BC catchment). The x-axis shows forecast CRPS values and the y-axis shows the ensemble size. Theoretical optimums correspond to results of subsets selected by the GA method, for different ensemble sizes. The Full-ESP corresponds to results of considering all historical scenarios, as is usually implemented in ESP studies. CT-Q and CT-P correspond to the CT method, considering the streamflow and precipitation as predictor variables, respectively. 'CT-Q Perfect' and 'CT-P Perfect' correspond to results of CT-Q and CT-P while having access to the actual state of the forecast period. The analogue markers identify the CRPS skill using the Analogue method considering streamflow as the predictor variable.

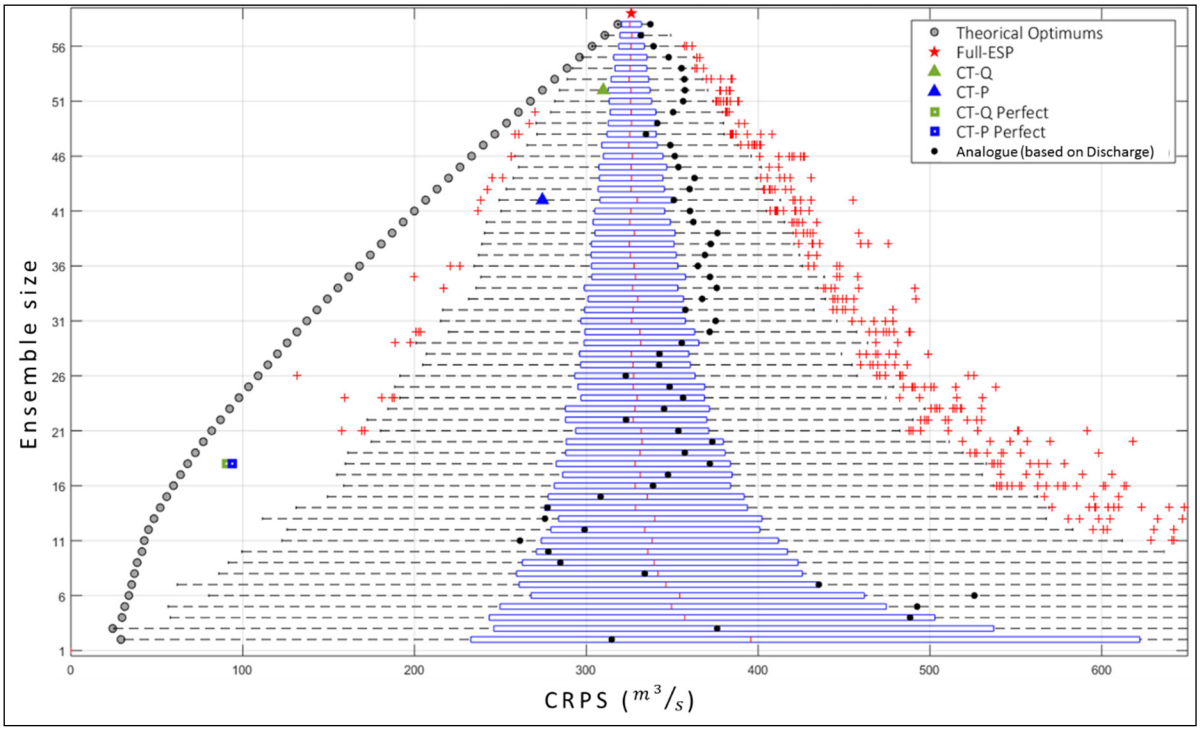


Figure 3.3 CRPS values for different combinations of members in the ESP forecast as a function the ensemble size for the QC catchment. The boxplots represent the values obtained with the 1000 random samplings for the July 1st, 1990, forecast date.

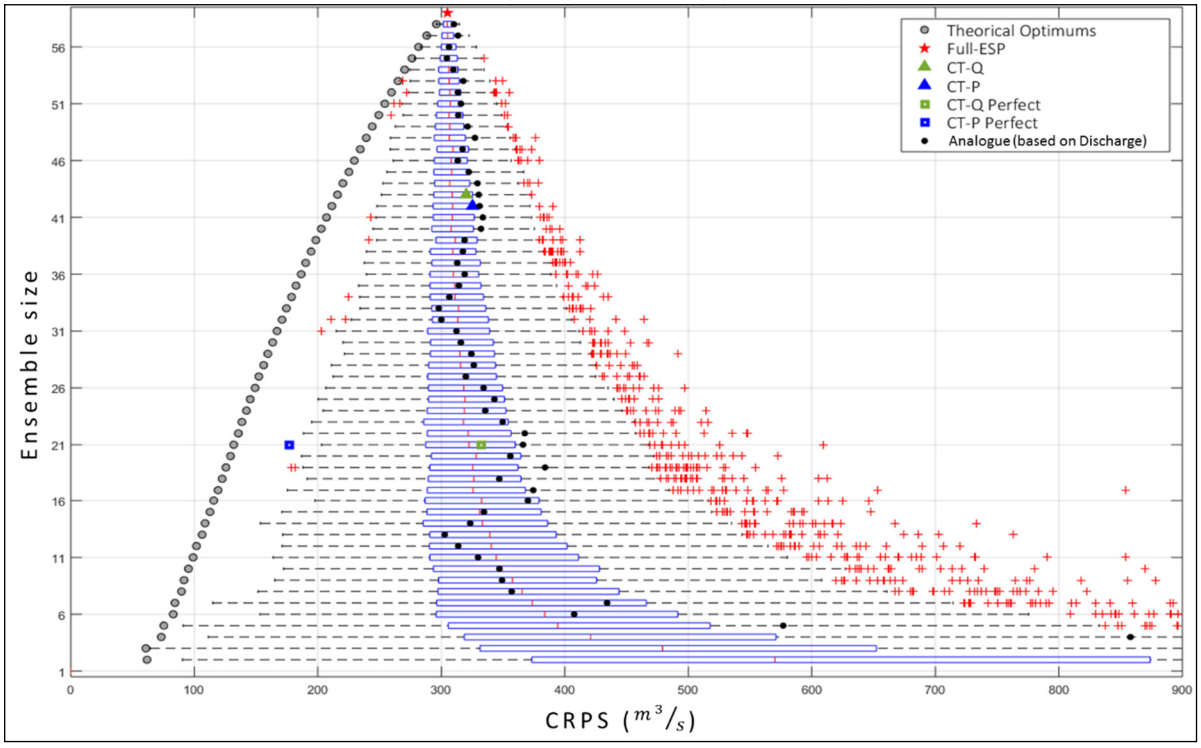


Figure 3.4 CRPS values for different combinations of members in the ESP forecast as a function the ensemble size for the BC catchment. The boxplots represent the values obtained with the 1000 random samplings for the July 1st, 1990. forecast date.

Multiple interesting patterns emerge from Figure 3.3 and Figure 3.4. First, random sampling either does not improve (in the best case), or indeed deteriorates (in the worse case) the CRPS versus using the whole ensemble. The fewer the members, the worse the median value. The pattern seen in Figure 3.3 and Figure 3.4 was reproduced in essentially all tests performed, i.e. for each of the 12 initial forecast start dates matching the 1st of each month) and various lengths of predictor and forecast windows (not shown). This observation is in agreement with the findings of Ferro et al. (2008). Some member combinations perform better than others, but without any other discriminating information, adding more members seems to be the correct approach. It can also be seen that the best combinations found by the GA algorithm (the theoretical optimal points) progressively improve as the number of members decreases, except

for the single member case. This indicates that the information is present in the ESP ensemble, but we must find a way to extract it adequately. This can be seen as the potential improvement for the forecast date, i.e., the maximum improvement that could be possible to gain from selecting members perfectly. As for the conditional methods (CT) and their variants, both with Discharge (Q) and precipitation (P), the results vary by catchment. The CT using available information performs better than Full-ESP in the QC catchment, but the opposite was observed in the BC catchment. Comparing CT results with and without having access to perfect information shows that merely having access to this information does not automatically improve the results. This could be due to the fact that CT is probabilistic, and therefore contains a certain amount of uncertainty. Hence, knowing the future state will not automatically mean that the ensemble members that are selected will perform better. A more thorough investigation is presented in section 4.1. The analogue method performs similarly, if not slightly worse, to that of the Full-ESP, albeit with a certain variability based on the ensemble size. None of the tested methods come close to systematically matching the GA-optimal forecast skill. This may be seen as an inability of the proposed methods to properly describe the uncertainty in the forecasting process and in the future forecast state. In other words, the factors that have been used to help identify future forecast states in the proposed methods do not provide enough insight to improve upon the Full-ESP in a reliable way.

These results have to be interpreted carefully as they represent a single snapshot of the forecast performance for a single day, and cover a single set of predictor and forecast periods. The process was then repeated for each year in the dataset, providing results for the years 1955 to 2004 for the BC catchment, as shown in Figure 3.5.

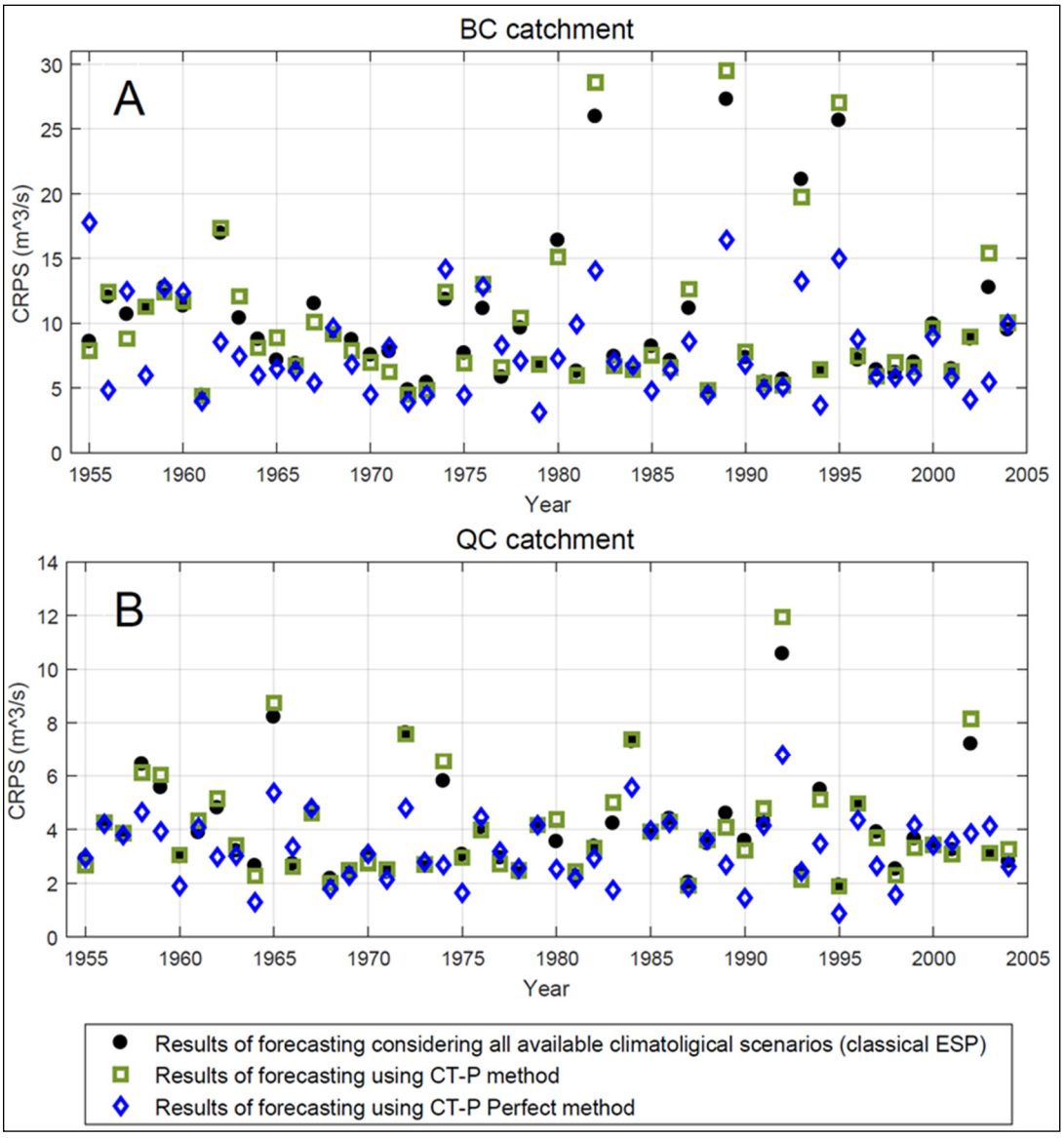


Figure 3.5 CRPS value for three ESP-based methods: (1) Full-ESP, (2) CT using precipitation as a predictor and (3) CT with precipitation and with perfect knowledge of the future state for the BC catchment. The predictor and forward periods are 90 days each for all July 1st forecast dates in the dataset from 1955 to 2004 inclusively

Figure 3.5 shows CRPS values for the 50 initial dates of forecast starting on July 1st of each available year for three ESP sampling methods, using a 90-day predictor period, and a forward period of the same duration for the BC catchment (A) and QC catchment (B). In Figure 3.5, the Full-ESP method is compared to (1) the CT with precipitation as a predictor variable and

(2) the CT with precipitation as a predictor but with perfect knowledge of the future precipitation classification. For the BC catchment, in 37 cases out of 50, the perfect information method scored the lowest CRPS values, indicating its better forecast skill versus the other methods. In the remaining cases, the Full-ESP method performed slightly better than the perfect information method. Results are similar for the QC catchment. The 74% success rate seen with the perfect information method here may be the result of knowledge of the true future state in a categorical sense (i.e., knowing if it will be wetter or dryer than average), without actually specifying the precipitation amounts. In other words, improving long-term precipitation forecast trends could significantly improve the ESP performance. However, since the method is not systematically better, even in the case of perfect knowledge, there would still be some untapped potential that could perhaps be accessed by using other variables and predictors. It is important to note that the results shown in Figure 3.3, Figure 3.4, and Figure 3.5 point to the Full-ESP method being similar or slightly better than the analogue and CT methods, depending on the number of ensemble members. A series of tests on other forecast periods, lead-times and predictor period lengths showed similar results, and the CRPS was typically better for Full-ESP than for reduced-size ensembles implementing the analogue or CT methods (results not shown).

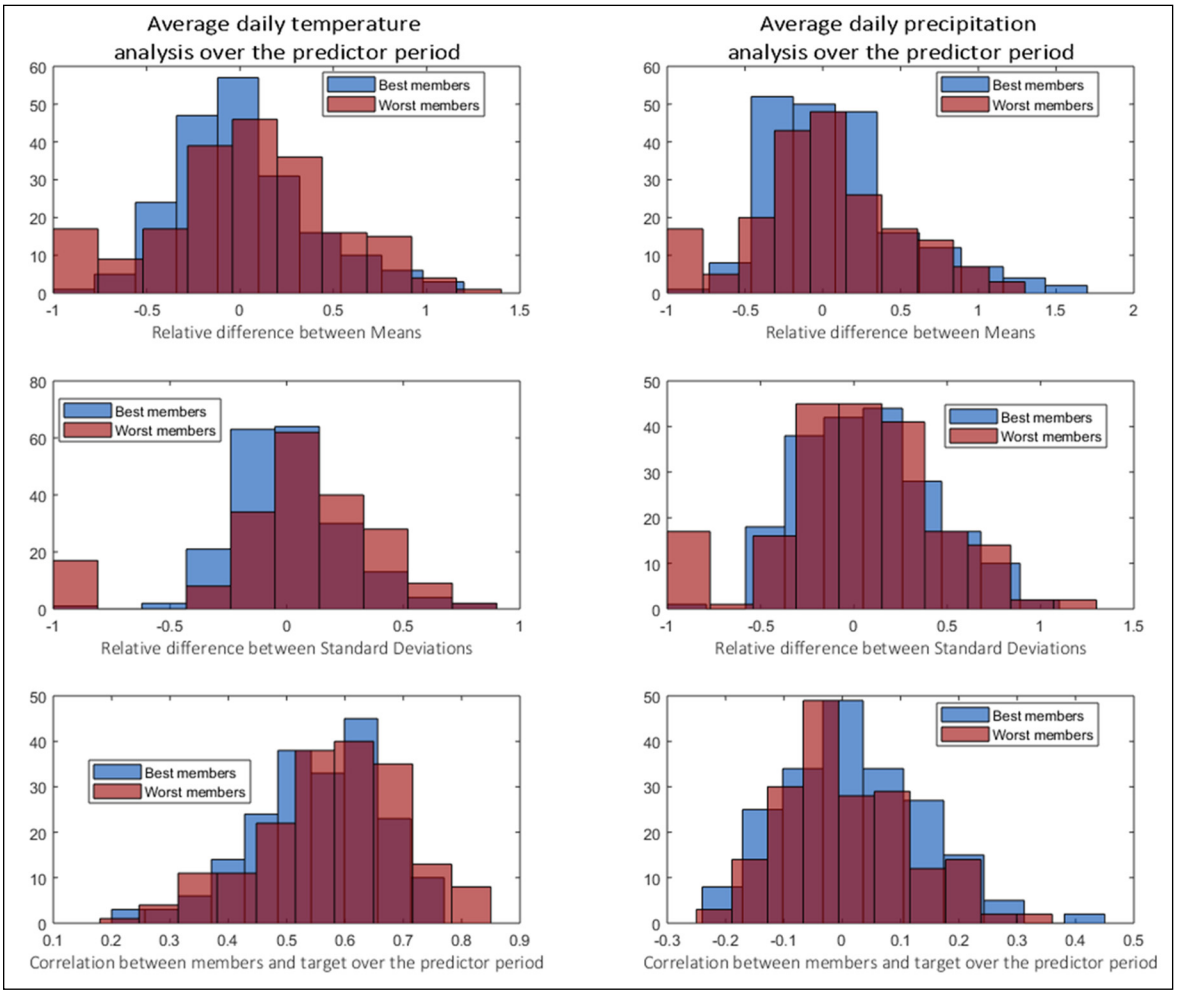


Figure 3.6 Comparison between desirable and undesirable members, as selected by the GA method, with respect to their hydrometeorological properties in the predictor period for the BC catchment. Left-hand histograms correspond to temperature data, and right-hand histograms represent precipitation data.

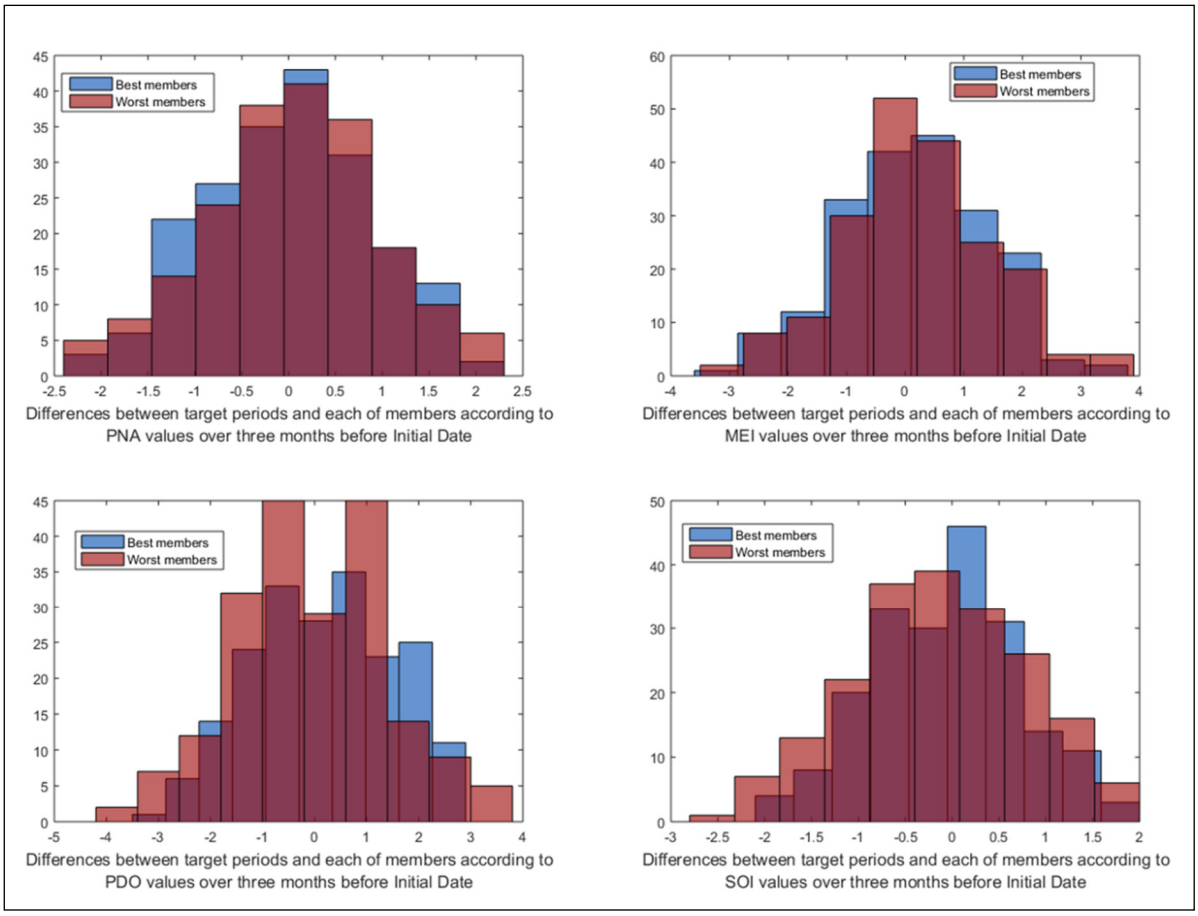


Figure 3.7 Comparison between desirable and undesirable members, as selected by the GA method, with respect to their large-scale climate indices in the predictor period for the BC catchment. Left-hand histograms correspond to temperature data, and right-hand histograms represent precipitation data. Climate indices represent the Pacific-North American (PNA); Multivariate El Niño/Southern Oscillation Index (MEI); Pacific Decadal Oscillation (PDO) and Southern Oscillation Index (SOI).

Figure 3.6 and Figure 3.7 show the results of the analysis of the ensemble member characteristics on forecasting performance for the BC catchment, for the four previously defined indices. These analyses were done for July 1$^{st}$ as the forecast issue date and were repeated 40 times (from 1966 to 2005), considering 90 days before the forecast date, and 90 days following the forecast date. Results are similar for the QC catchment, which are not shown.

In Figure 3.6 , the histograms each include 200 values, i.e., 40 years x 5 members per year. The x-axis represents the relative difference between the predictor period of the selected ESP member and the forward period. The predictive variables are the temperature (left column) and precipitation (right column), and the three rows show the results when comparing the difference in means (top row), standard deviations (center row) and correlations (bottom row). The blue and red histograms show the results for the best and worst members, respectively. It can be seen that the distributions of both histograms are very similar, and that none of the selected variables allowed generating a clear pathway to separate good members from bad ones. Figure 3.7 shows the distributions of the relative difference between (1) the averages of climate indices in the predictor period of each of five best (worst) members for the 40 forecast years, and (2) the values of the associated forward period indices.

The results show that while there seems to be some minor differences in some of the distributions, especially in the variability of the PDO and SOI distributions, these differences are not statistically significant, and do not allow for efficient discrimination between the desirable and undesirable members. The analyses in Figure 3.6 and Figure 3.7 were performed on a multitude of forecast dates, predictor period durations and forecast lead-times with similar results, which are not shown. These predictor variables thus seem insufficient to allow the identification of members that maximize the forecast skill.

## 3.5 Analysis and discussion

In this study, we proposed a procedure to quantify the theoretical potential of using historical weather observations as probable scenarios in an ESP approach for long-term discharge forecasting. Typically, all members in ESP forecasting are considered as equally probable. However, in this study, we used a procedure to determine the best possible combinations of members for different ESP ensemble sizes. The GA approach allowed finding these combinations in hindcasting mode, with perfect knowledge about the hydrometeorological states to be forecasted. While this approach is not available to the forecaster in real time, the

rationale, however, was to use it to quantify the theoretical limit that could be achieved using the historical climatology on hand.

Various ESP member selection techniques were tested to analyze their ranking as compared to the theoretical optimum and the Full-ESP ensemble. Each method was tested with a range of parameters, such as the initial date of forecast, the duration of the predictor period, and the lead-time. In almost all tests, lower CRPS values could be obtained with smaller ensemble sizes, but at the same time, expected CRPS values increased if the members of the subset were selected randomly.

One important message obtained from this study is that when performing an ESP forecast using climatology, there is a strong chance that some members present in the dataset could possibly be excluded in order to improve the forecast skill. In that sense, decreasing the ensemble size can improve the forecast skill through the elimination of improbable scenarios. This selection also carries a risk because a desirable member may be removed instead of an undesirable one, in which case the forecast skill will degrade. Therefore, research efforts should focus on increasing the ability to probabilistically categorize the future hydrological regime (wetter or dryer than average) and condition the ESP member selection on those probabilities. As noted previously, many studies have investigated this very subject, either through climate indices, long-term NWP forecasts, or other means, and the present study will help quantify the potential skill attainable with those methods.

As shown in Figure 3.3 and Figure 3.4, and from the numerous results obtained during the tests performed in this study, the results of the CT and the analogue methods were highly variable, and depended on the lead-time, the predictor period length, the predictor variable, the initial date of forecast and the catchment characteristics. In the case of CT, the method was not reliable enough to be recommended for use when the predictor was the streamflow or the precipitation.

CT-Perfect variants were implemented to better test the potential of the CT method. As such, it was conducted under conditions without uncertainty regarding precipitation during the forecast period. On average, CT-Perfect results were much better than those obtained using the Full-ESP, meaning that improving the overall forecasted precipitation category (wet or dry) accounts for a significant portion of the ability of the methods to select desirable members. The skill could probably be further increased by finding better predictors than the average previous precipitation or streamflow. For example, in a multivariate setting, adding information from climate indices and combinations thereof to the precipitation and temperature could be envisioned; as well, consideration could also be given to using other statistics than those used in this study. The analogue method was generally less skillful than the Full-ESP, and this could be attributed to the same problem encountered with the CT method, i.e., the predictor period variables are not good enough predictors for future inflows or precipitation. Although the results of the member selection methods shown herein did lead to forecast skill improvements, the methods presented could be useful to researchers in ESP forecasting and water resource management, for two reasons. First, b*efore implementing an ESP forecasting system, the* procedure could be used in hindcasting mode to determine how much potential the forecast may yield, given its specific parameters (initial date, predictor period duration and lead-time, available observation record). This could serve as a reference to help manage any expectations and determine the limits of the approach before any more efforts are deployed. Second, the procedure could be used to benchmark, analyze and improve existing methods that are currently being used in practice to generate ESP forecasts. This would allow quantifying improvements on a relative scale, with a known baseline value. This type of evaluation could be repeated as many times as needed in order to obtain a more comprehensive view of the method's performance.

The proposed method to investigate relationships and correlations between desirable members and hydroclimatic indicators could also be used to help develop member selection methods. In the general tests performed in this study, no strong relationship was found. In other words, simply considering the average value of precipitation or temperature during the predictor

period was not enough to indicate the forward period characteristics. More complex predictors could potentially allow such a relationship to be exposed. There are many other metrics and combinations of metrics that could be investigated, including, for instance those in Feng et al. (2016), Hossain et al. (2018), Hossain et al. (2020) and Konapala et al. (2018). Non-linear combination methods could also be tested with this approach, which was out of the scope of this paper.

This study proposes a method to quantify the skill of an ESP forecast using reduced ensemble sizes by selecting fewer, but more desirable members. However, there are some general limitations to using this approach. First, the historical dataset must be long enough to represent multiple climatological states over the catchment in order to have a realistic chance of portraying the climatological uncertainty. Selecting a subset of an already small ensemble would prove difficult as the transition probabilities would be ill-defined. Second, the fact that the ensemble size changes during the subsetting process induces a bias in the CRPS calculation, as shown in (Ferro et al., 2008). This effect was not explicitly accounted for in this study, but the impact should be minimal because (1) the response is similar to what would be expected with the corrections and (2) even with our best methods and attempts, we were unsuccessful at identifying which of the members in the reduced ensemble should be selected. Therefore, even for a given size ensemble we were unable to improve the method. Also, climate change signals could be present in the observed dataset used to define the ESP members. There is a legitimate claim that ESP members from multiple decades ago could be unrepresentative of the current climatological state. In such cases, while care should be taken when building the ESP ensemble, determining how to do so is a challenge in itself. It is possible that by weighting the members differently – or by excluding some entirely – a more reliable ensemble could be built. Another limitation relates to the selection of catchments used in this study. The two catchments are snowmelt dominated and thus have somewhat similar hydrological regimes. Perhaps this method could provide different results in contrasting climate regimes, although the identification of the maximum potential skill would probably be similar. Finally, it is important to note that this study did not make use of pre- or post-processing of ensemble members, but

rather, used historical data directly. The pre-processing step could conceivably use the quantification and selection methods in this study to improve the accuracy of processed ensembles for long-term streamflow forecasting.

## 3.6 Conclusion

This study presents a procedure to quantify the potential skill of ESP using historical climatological data. In typical implementations of ESP, all available members are used to add diversity to the ensemble, and all members are considered equally likely to occur in the future. However, knowing some information about the future state can help focus the ESP member selection on the expected conditions. One way of doing so is by removing members that are unlikely to be realized in the future. This study describes a method to determine the best possible members for any ensemble size and to weigh the risk and potential benefit of using a member selection approach. Three approaches were tested, and while two (CT and analogue) did not lead to skill improvement, the CT-Perfect procedure allowed to determine that classifying the climatological state of the forecast period could improve long-term forecasting accuracy.

The maximum potential improvement was quantified using a GA method to find the best combination of members for any ensemble size. It provides a means of comparison between results of various methods, such as the random selection of members and the theoretical optimum to the Full-ESP ensemble. This quantification could help improve ESP forecasting by providing a more consistent baseline than relative reductions in CRPS, as is typically done in the literature.

Finally, a method was proposed to help differentiate desirable from the undesirable members, in a consistent manner, for ESP forecasting using the theoretically optimal ESP members' properties and correlating them to hydroclimatological indices in hindcasting mode. Although the results show that the selected variables are not good predictors of this relationship, the

method could be employed in future research to attempt to determine the most efficient predictors and quantify their levels of success.

# CHAPTER 4

# SENSITIVITY ANALYSIS OF THE HYPERPARAMETERS OF AN ENSEMBLE KALMAN FILTER APPLICATION ON A SEMI-DISTRIBUTED HYDROLOGICAL MODEL FOR STREAMFLOW FORECASTING

Behmard Sabzipour [a], Richard Arsenault [a], Magali Troin [b], Jean-Luc Martel [a], François Brissette [a]

[a] Hydrology, Climate and Climate Change Laboratory, École de technologie supérieure, Montréal, Canada.
[b] HydroClimat| TVT, Maison du Numérique et de l'Innovation,Toulon, France

Paper submitted to *the Journal of Hydrology*, May 2022

## Abstract

Data Assimilation (DA) is an important step for improving prediction accuracy and real-time correction of hydrological models for operational forecasting purposes. The Ensemble Kalman filter (EnKF) is one of the most popular techniques used to address the issues of updating the model states and parameters by creating a novel set of initial conditions in real-time. This study aims at identifying optimal seasonal EnKF parametrizations to reduce the uncertainty associated with the initial conditions in a semi-distributed hydrological model of a snow-dominated catchment in Canada. Sensitivity analysis is performed to evaluate the effects of the EnKF individual hyperparameters (temperature, precipitation and inflow uncertainties) and the updating of the water content of three state variables (vadose zone, saturated zone and snowpack) on the skill of short-term (up to 9 days lead-time) forecasts. Results show that the performance of the forecasts is sensitive to the individual hyperparameters and particularly so to the temperature uncertainty, which varies between seasons. Additionally, the forecast skill is related to the choice of the state variables to be updated depending on the season. The vadose zone state variable displays higher importance and sensitivity than the other states, and the

results indicate that all state variables should not be systematically updated. Finally, combining the best hyperparameter values with the optimal combination of state variables to update, insight is provided on the success of the DA scheme that is evaluated on a rolling horizon using a set of seasonal rules to gain forecast performance over the span of multiple years.

**Highlights**

- Reducing initial forecasting errors using the EnKF DA technique.
- Improving real-time streamflow forecasts by an optimal EnKF implementation.
- A seasonal adjustment of the EnKF hyperparameters improves forecast accuracy.
- The vadose zone state variable is the most impactful and should be prioritized.
- Data assimilation is most sensitive to temperature uncertainty for forecasting

**Keywords:**
Data Assimilation; Sensitivity Analysis; Ensemble Kalman filter; Ensemble Streamflow Prediction; CRPS.

## 4.1 Introduction

Characterizing and communicating predictive uncertainty in hydrological forecasting is critical to effectively support water planning and management decisions. In recent years, much progress has been made in forecasting research, and additional efforts introduced towards uncertainty quantification and reduction in hydrological forecasting (Roundy et al., 2019). Uncertainty in hydrological forecasts originates from three sources: 1) initial conditions of the atmosphere derived from observations and numerical weather forecasts (Sun et al., 2018), 2) numerical weather prediction (NWP) model uncertainty (Zappa et al., 2010), and 3) hydrological model uncertainty (Bourgin et al., 2014). Hydrological predictive uncertainty includes errors in model structure, parameters, initial conditions, and calibration (Ajami et al., 2007; Li et al., 2009). Over the last two decades, DA has attracted attention as a relevant method for increasing the skill of hydrological forecasts by improving initial model states,

model parameters and structures, while accounting for all predictive uncertainties (Abbaszadeh et al., 2018; DeChant & Moradkhani, 2014; Liu et al., 2012).

The purpose of DA is to integrate various observations (i.e., weather forcings, streamflow, soil moisture, snow cover), which allows the updating of the model state in time for keeping the forecasted streamflow in line with their observed counterparts at the moment of producing a forecast. DA ranges from simple rule-based, direct insertion methods to more advanced methods (Abbasnezhadi et al., 2021; Liu et al., 2012). The advanced DA methods are usually classified according to their tractability approximations: linear (e.g., Kalman filter (KF)) versus nonlinear (e.g., EKF), filters (e.g., PF) versus smoothers (e.g., variational DA method), and deterministic (e.g., KF and EKF) versus ensemble (e.g., EnKF DA) (Moradkhani et al., 2018). Among the popular advanced DA methods listed in Troin et al. (2021), the KF approach is one of the most widely used methods, because of its efficiency, simplicity of application, flexibility for coupling with hydrological models and low computation demand (Sun et al., 2016). The KF variants include, for instance, the EKF, the EnKF, the Asynchronous EnKF (AEnKF) and REnKF. The EKF relies on a Taylor extension scheme for linearizing the nonlinear system, while the ensemble KF techniques avoid direct linearization by statistically analyzing the ensemble members (Sun et al., 2016). KF techniques outperform other DA approaches for distributed hydrological models showing the advantage of the KF approaches in updating distributed states through error covariance modelling (Abaza et al., 2017; Khaki et al., 2017; Mazzoleni et al., 2018; Moradkhani et al., 2018; Thiboult et al., 2016). Even though they increase the computational cost, the ensemble KF techniques have the advantage of explicitly treating uncertainties due to the ensemble nature with equal weighting of each member (Abaza et al., 2017; Khaki et al., 2017; Mazzoleni et al., 2018; Moradkhani et al., 2018; Thiboult et al., 2016). Therefore, the EnKF still remains the most widely used hydrological DA method.

The EnKF techniques enhance forecasting skill, however, some challenges still remain regarding their effective implementation: this includes the aspects of joint parameter estimation, state/error updating, timing errors, multi-observational and large-scale DA (Sun et

al., 2016; Troin et al., 2021). The performance of the EnKF DA procedure depends on the specification of the hyperparameters (input/output perturbations) and the selection of updated state variables (Moradkhani et al., 2005), as not all state variables can (or should) be updated (Thiboult & Anctil, 2015; Thiboult et al., 2016). However, most studies rely on an arbitrary selection of the hyperparameters and updated states for the EnKF implementation, which are mostly defined on an annual basis (Thiboult et al., 2016), without accounting for the importance of seasonal effects for an operational streamflow forecasting system. The calibration of the hyperparameters is also an important issue where the calibration metric can influence the reliability of streamflow forecasts (Bergeron et al., 2021). Several studies show that EnKF provides ensemble streamflow forecasts with higher performance, skill and reliability for real time and short lead-times compared to longer lead-times (Reichle, 2008; Thiboult et al., 2020; Vergara et al., 2014). Nevertheless, in terms of efficiency and temporal persistence of the updating effect, EnKF-based forecasts appear to provide more benefit compared to other empirical (Jiménez et al., 2019) and sequential ensemble-based DA methods (Piazzi et al., 2021) for increasing lead-times.

Considering the above-mentioned challenges regarding the application of EnKF, this study aims at identifying the optimal EnKF parametrization at a seasonal scale to quantify and reduce the predictive uncertainties related to initial conditions (state variables) in an operational forecast mode. The seasonal scale was selected as a compromise between the yearly scale (which could not pick up intra-annual changes in the hydrological processes) and the monthly scale (which would have required more data to evaluate correctly). A second objective of the study is to assess the EnKF performance regarding the efficiency of the season-based updating rules for increasing lead-times. The experiment is conducted through the CEQUEAU semi-distributed hydrological model (Morin & Paquet, 2007) over the Lac-Saint-Jean catchment in Canada. The catchment and the hydropower system it contains are operated by Rio Tinto Power operations, a subsidiary of Rio Tinto, one of the largest mining and metals companies in the world that uses this hydropower system to generate electricity required in their aluminum smelters in the region. Section 4.2 presents the experimental design of the study. Section 4.3

depicts the key results of the sensitivity analysis related to the EnKF parametrization followed by a discussion in Section 4.4. Concluding statements are provided in Section 4.5.

## 4.2 Materials and methods

### 4.2.1 Study site

The case study is applied to the Lac-Saint-Jean catchment in central Quebec, Canada, which contains a large hydropower reservoir managed by Rio Tinto to provide energy for their aluminum smelters (see Figure 4.1). The drainage area is 45 000 $km^2$ and the hydrological regime is strongly influenced by snow accumulation and melt (Sabzipour et al. 2021). Rio Tinto must therefore plan water resources management under varying conditions throughout the year, including more reactive summer and fall seasons, the long low-flow conditions of winter and the large inflows caused by snowmelt during the spring. Furthermore, precipitation falls mostly in summer (336 mm on average over the months of June, July and August) and fall (287 mm over September, October and November). Winter and spring see lower precipitation amounts (239 mm over December, January, February and March for winter, and 145 mm for April and May in spring). The reservoir has an area of approximately 1000 $km^2$ and can store up to 4550 $hm^3$ of water (Arsenault & Côté, 2019). The generating station can turbine 1600 $m^3/s$ and spillways can release as much as 6000 $m^3/s$ during flood events, depending on the reservoir hydraulic head. This gives flexibility to the water resources managers, but optimizing the water usage requires high-quality hydrological forecasts over various lead-times. Rio Tinto therefore uses the CEQUEAU semi-distributed hydrological model driven by operational NWP forecasts as inputs for short-term hydrological forecasting and then adding climatological weather to extend the forecasts beyond 10 days for longer-term forecasts. These forecasts depend on the initial states of the hydrological model, which need to reflect the initial conditions of the catchment prior to issuing the hydrological forecast. Therefore, a DA process is typically performed manually by Rio Tinto's hydrologists. This leads to a problem of repeatability for studying the forecasting skill over longer periods, which

paves the way for this study, where automatic DA is implemented on the semi-distributed CEQUEAU model over the Lac-Saint-Jean catchment. This study therefore requires various types of data, as described in the following section.

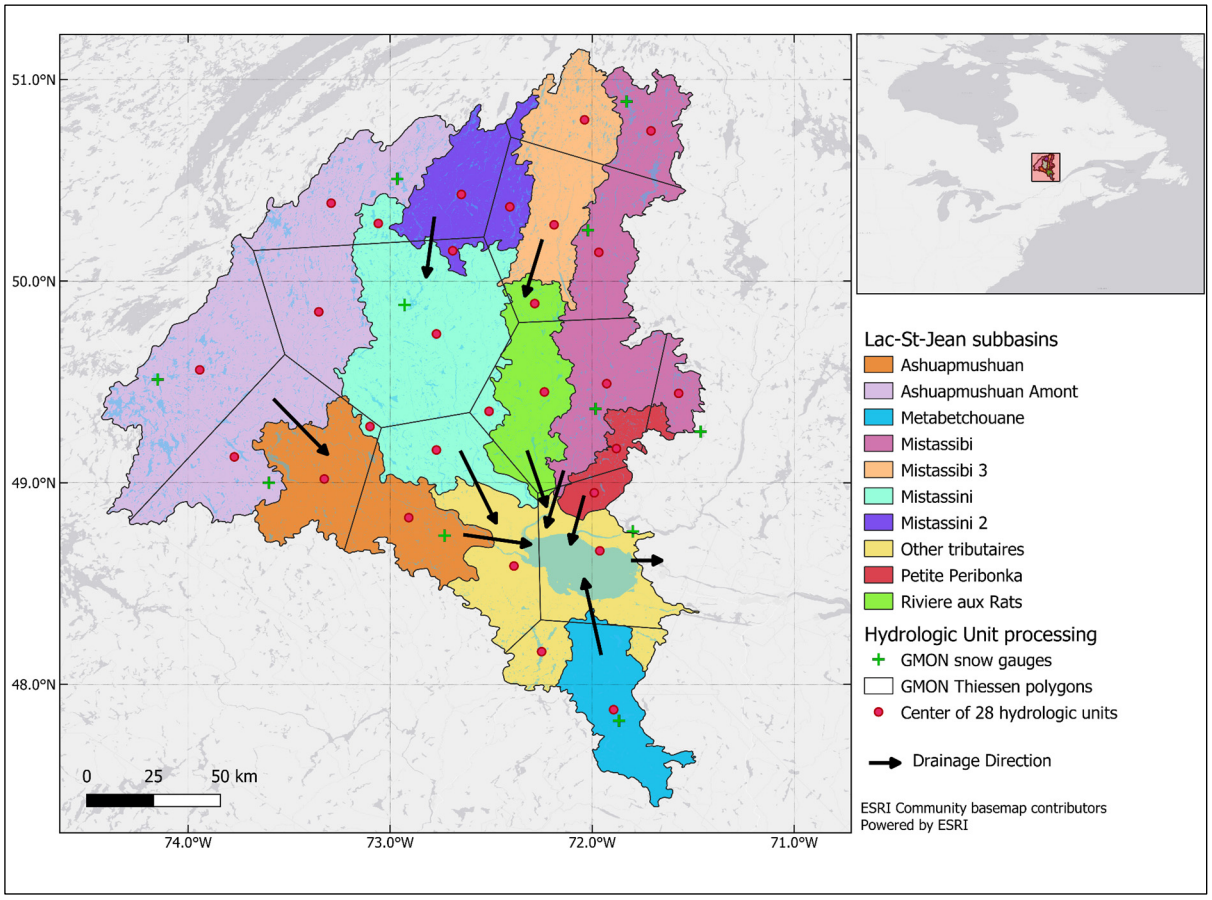


Figure 4.1 The Lac St-Jean (LSJ) catchment and its ten contributing sub-catchments, including the "other tributaries" which are the sum of all smaller ungauged rivers flowing into the LSJ reservoir. The eleven gamma ray snow sensors (GMON) are represented by green crosses, and the Thiessen polygons they generate are also shown. The intersection of the ten sub-catchments and the eleven Thiessen polygons represent the 28 hydrological units of the CEQUEAU semi-distributed hydrological model used in this study and are represented by red circles. Arrows show the drainage direction between the sub-catchments as well as the location of the basin outlet.

#### 4.2.1.1 Datasets

Three types of data are required to perform this study: observed inflows at the hydropower reservoir, observed meteorological data for model simulation and DA, and forecast data from a NWP model to provide the weather forecasts. Inflow data is provided by Rio Tinto and covers the 1954-2019 period inclusively. The daily inflow data is computed through mass balance, using measured reservoir levels and known water releases to estimate the actual inflows to the reservoir. Evaporation is not directly estimated but is included in the daily water level fluctuations, which can lead to some hot summer days having small underestimations of inflows. This method can lead to noisy inflow patterns, therefore a 3-day moving window average is performed to smooth out inconsistencies caused by wind, waves or other small temporary changes in water levels during inflow computation.

Meteorological data is also provided by Rio Tinto for the same period as the inflows (1954-2019). This data is a spatially interpolated grid from Rio Tinto's 20 weather stations on the catchment, as well as approximately 18 stations positioned outside the catchment boundaries. The result is a 48-cell grid of precipitation and temperatures over the catchment for the entire period. Daily meteorological values include total precipitation and average daily temperature, which are the only meteorological inputs required by CEQUEAU.

Finally, the operational weather forecast data are provided by the European Center for Medium-range Weather Forecasts (ECMWF). 6-hourly precipitation, minimum and maximum temperatures are downloaded for the period ranging between May 20, 2015, and December 31, 2019. Prior data are not considered since the ECMWF Integrated Forecasting System (IFS) was upgraded considerably in May 2015, which would have changed the forecast statistics significantly. Total daily precipitation data are computed from the 6-hourly periods, as is the average daily temperature. The ECMWF forecasts used in this study are the 50-member perturbed forecasts for 1 to 9 days of lead-time to reflect the operational implementation of

these forecasts at Rio Tinto. The data are available from the ECMWF MARS archive on the ECMWF website: https://www.ecmwf.int/en/forecasts/dataset/operational-archive.

### 4.2.2 Methodology

This section details the methodology used to generate ensemble streamflow forecasts (Figure 2). The methodology can be broken down into three parts: 1) the generation of an ensemble of initial hydrological model states through perturbed meteorological inputs, 2) the DA step where the initial conditions are optimized, and 3) the forecasting step using the assimilated initial conditions. The process then evolves for the next day, allowing to evaluate the method over a period of 4.5 years. The forecast skill is then compared to that obtained in an open loop run where no DA is performed. The data assimilation loop is presented in figure 2. The open-loop forecast is not shown as it is the classical approach without any data assimilation, using previous model states to perform the next step's forecast.

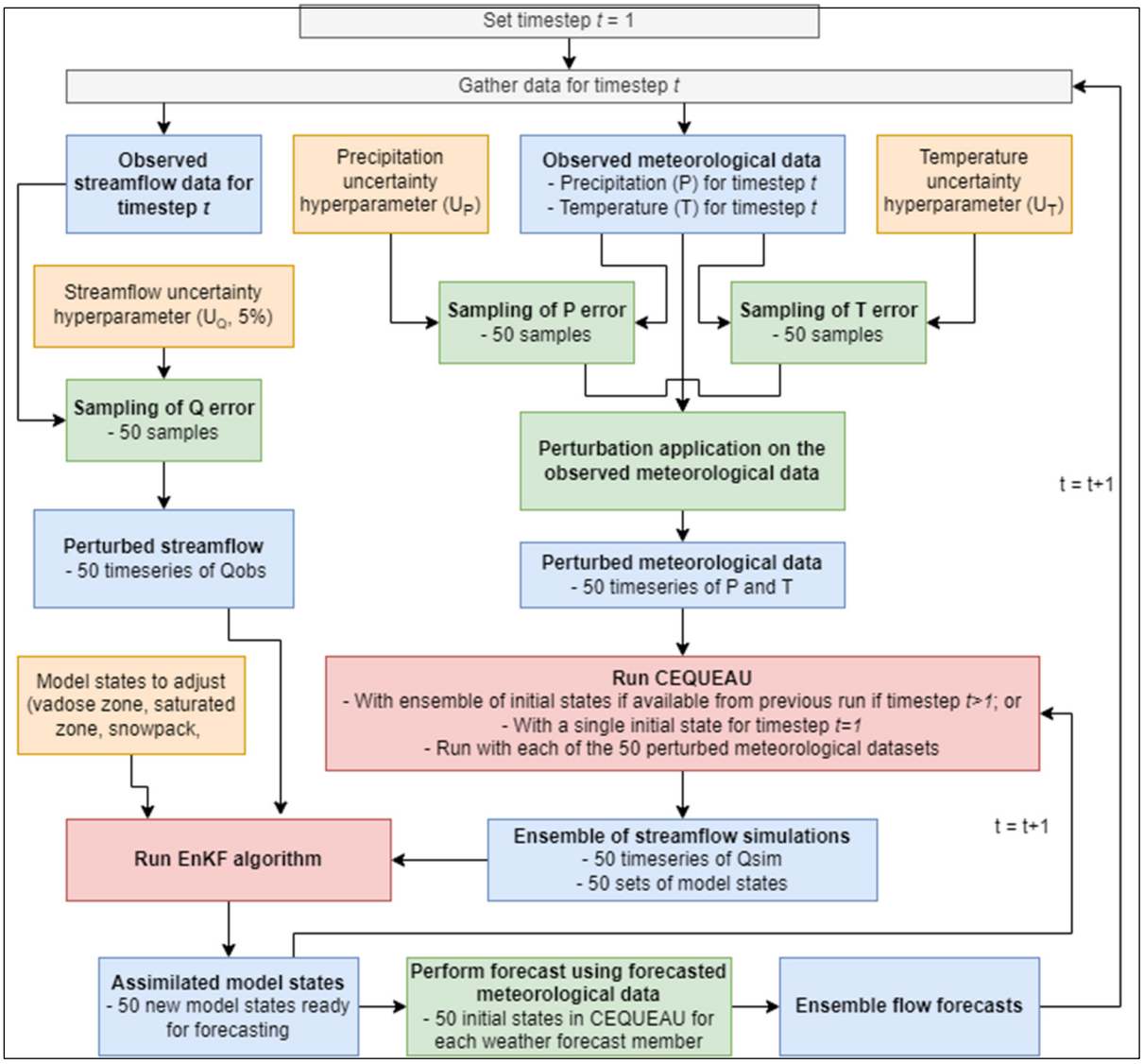


Figure 4.2 Flowchart demonstrating the preparation of perturbed weather data, the assimilation process and the generation of flow forecasts. Weather forecasts used in this study are the operational ECMWF ensemble (perturbed) forecasts over the historical period. The process shown in the flowchart is repeated for every combination of $U_P$, $U_T$ and combination of model states to adjust.

#### 4.2.2.1 Hydrological model

The CEQUEAU hydrological model is used in this study (Morin & Paquet, 2007). CEQUEAU is a semi-distributed model with three reservoirs to represent key hydrologic processes: a surface reservoir to simulate near-surface processes (i.e., evapotranspiration and infiltration), a deeper reservoir to simulate groundwater and base flows, and the ability to accumulate and melt snow (snowpack modelling). Evapotranspiration is estimated with the Oudin method (Oudin et al., 2005) and snow processes are simulated with the degree-day CEMANEIGE model (Valéry, 2010). The watershed was divided into 28 hydrological units, corresponding to the intersection of the 10 sub-catchments of approximately similar hydrological conditions (for semi-distributed hydrological modeling) shown in Figure 4.1 and the region of influence of snow sensor (gamma ray monitors; GMON) stations. The 28 hydrological units are also shown in Figure 4.1, where red circles identify the hydrological unit centroids. Each hydrological unit receives uniform precipitation and computes vertical fluxes independently before being routed to the outlet. CEQUEAU only requires daily precipitation and daily average temperature as meteorological inputs, and 23 parameters need to be calibrated. The calibration of the CEQUEAU model was performed by Rio Tinto, who provided the pre-calibrated operational model of the entire system for this study.

#### 4.2.2.2 EnKF

The DA method used in this study is the EnKF, which is a Monte Carlo approach dedicated for nonlinear filtering issues (Bergeron et al., 2021; Evensen, 1994; Piazzi et al., 2021). The EnKF is extensively used in hydrological sciences and forecasting applications (Bergeron et al., 2021; Evensen, 1994; Piazzi et al., 2021). The EnKF relies on the approximation of the conditional probability densities of the state error covariance by a finite large number of randomly generated model realizations. It produces an ensemble of members which are

independently propagated through the model operator. *A priori* state error covariance is derived from the statistical analysis of the ensemble (Evensen, 1994).

The two phases of EnKF are the following: (1) a prior estimation - *a priori* state ensemble is first created by adding some random perturbations to the best estimated initial model state. The matrices of a prior error covariance are estimated from the state and model output ensemble error matrices; (2) a posterior estimation - the state ensemble is integrated forward in parallel based on the original model. The observations are perturbed by adding Gaussian (i.e., for temperature and streamflow) or Gamma (i.e., for precipitation) noise from the error distribution (sampling errors). Then, the *a posteriori* error covariance matrix is estimated from the ensemble error matrix of the *a posteriori* states. The Kalman gain is then computed from the forecast error covariance matrices. During this recursive process, EnKF tries to minimize the differences between the estimated model state (i.e., the simulated streamflow ensemble) and the observed streamflow (which also includes uncertainty) (Clark et al., 2008; Liu & Gupta, 2007; Moradkhani et al., 2005).

#### 4.2.2.3 DA experiment

The experiments considered in this study are based on the assimilation of streamflow at the catchment outlet into the CEQUEAU model. The details related to the procedure used to assimilate the observations are presented hereafter.

The CEQUEAU model uncertainties are evaluated by the EnKF through the probabilistic distribution of the ensemble members of simulated streamflow. Observed meteorological forcing data (temperature and precipitation) are perturbed to provide a distribution of inputs for each day, which are then transformed into a distribution of simulated flows by CEQUEAU. Observed streamflow uncertainty is also accounted for by introducing an error sampled from a Gaussian distribution. All uncertainty quantities are defined by sampling values within probable bounds, testing various combinations and evaluating their impacts on forecast skill.

Equations (4.1) to (4.5) show how the perturbed meteorological and hydrological timeseries are computed.

$$T_{p,t} = N(T_t\ , U_T) \tag{4.1}$$

$$k_t = \frac{{P_t}^2}{\left(P_t \times\ U_p\right)^2} \tag{4.2}$$

$$\theta_t = \frac{\left(P_t\ \times\ U_p\right)^2}{P_t} \tag{4.3}$$

$$P_{p,t} = \Gamma(k_t, \theta_t) \tag{4.4}$$

$$Q_{p,t} = N\left(Q_t\ , Q_t\ \times U_Q\right) \tag{4.5}$$

Where:

- $T_{pert,t}$, $P_{pert,t}$ and $Q_{pert,t}$ are the perturbed temperature, precipitation and streamflow for timestep t, respectively.
- $T_t$, $P_t$ and $Q_t$ are the observed temperature, precipitation and streamflow for timestep t, respectively.
- $U_T$, $U_P$ and $U_Q$ are the uncertainty hyperparameters for temperature, precipitation and streamflow, respectively.
- $k_t$ and $\theta_t$ are the scale and shape parameters for the gamma distribution $\Gamma(\cdot)$ for day t;
- $N(\cdot)$ is the Normal distribution with parameters of mean and standard deviation; and
- t is the forecast day which covers all days from 1 to T, the total number of days for which a forecast is generated.

In this study, 25 samplings are taken for each variable, each day, and each sub-catchment to obtain an empirical distribution of the observed variables including the uncertainty as specified

by the respective hyperparameters. Special care is taken to ensure that negative streamflow values, which could theoretically occur during the perturbation process, are set to zero.

A one-year warm-up period is implemented to dissipate the initialisation errors. Ensembles are initialized with a perturbation of the state variables of the CEQUEAU model. Then, a first ensemble streamflow simulation of three days is produced by using the generated ensemble of perturbed temperature and precipitation, and the ensemble of model states. This allows the CEQUEAU model to generate an ensemble of simulated streamflows ($Q_{sim}$), which are compared to the ensemble of perturbed observed streamflows ($Q_{obs}$). The state variables of the CEQUEAU model are updated according to prior estimation of initial state ($Q_{sim}$) and posterior estimation ($Q_{obs}$ and the added error $Q_{obs\ error}$).
The selection of the best hyperparameter sets is conducted as follows:

1. Multiple independent DA runs are conducted from 2015-01-01 to 2019-06-04, by varying the uncertainty bounds of the precipitation and temperature inputs. The uncertainty values are modified by increments of 0.5°C from 0.5°C to 10°C, and the precipitation uncertainty values are modified by increments of 5% from 5% to 100%. Therefore, the search space is divided into [20 temperature x 20 precipitation = ] 400 possible uncertainty combinations, and DA is performed on the entire time series for each combination. It is important to note that these values (400 possible uncertainty combinations) represent the uncertainty added around the observed value, and thus the distribution is centered around the observation but has larger variance the higher the value of the hyperparameter. It also means that even if the uncertainty values are positive (e.g., 3°C), the errors sampled from the distribution can be negative as the uncertainty value refers to the standard deviation of the temperature adjustment distribution. Furthermore, multiple values of observed streamflow uncertainty were evaluated but the DA method was found to be only marginally sensitive to this uncertainty. Therefore, a single fixed value of 5% was selected for this part of the study,

to allow for numerical stability in the development of the observed streamflow distribution.

2. The three model state variables selected for the update are the contents of the vadose zone reservoir, the saturated zone reservoir and the snowpack. The states are varied for each of the sub-catchments independently, therefore when changing one state variable, a total of 28 cells are modified. The model state variables are updated both individually and in various combinations. This is done to first show the isolated effect of adjusting each state variable according to the season, and then to do the same in the context that multiple variables are allowed to change simultaneously. This is repeated for each of the 400 combinations as described in (1).
3. The hyperparameter set leading to the best forecast skill (as described in the next section) for each season (winter; DJFM, spring; AM, summer; JJA and fall; SON) is selected. An additional test is performed to identify if the combination of the state variables during state updating can improve the forecast skill. Thus, besides selecting the optimum hyperparameter set for each season, we also provide insights regarding which state variables need to be updated for each season.

#### 4.2.2.4 Performance evaluation

The performance of the different streamflow forecasts is assessed using the CRPS (Hersbach, 2000) at different lead-times (from day 1 to day 9) during the forecast. CRPS is a suitable metric for probabilistic forecasts. CRPS measures the average distance between the observed ($F(q_{obs})$), and the ensemble forecast ($F(q_{sim})$) probability density function, given by equation (4.6):

$$CRPS = \frac{1}{T}\sum_{t=1}^{T}\int_{-\infty}^{+\infty}\left[F(q_{sim_i}) - F(q_{obs})\right]^2 dx \tag{4.6}$$

where T is the total amount of time steps. The CRPS ranges from 0 to + ∞, zero being a perfect forecast. For streamflow, the CRPS unit is $m^3/s$.

Finally, the CRPSS (Continuous ranked probability skill score) is used to compare two forecasts:

$$CRPSS = 1 - \frac{CRPS_{forecast}}{CRPS_{reference}} \tag{4.7}$$

where $CRPS_{forecast}$ is the CRPS of the new tested forecast, and $CRPS_{reference}$ is the CRPS of the reference forecast, namely the open loop forecast in this study.

## 4.3 Results

### 4.3.1 Effects of meteorological data uncertainty on hydrological forecasting skill

This section provides insights on the hyperparameters defining observational uncertainty and how they affect the reliability of the ensemble forecasts. This is done by evaluating the effect of gradually increasing the individual temperature and precipitation uncertainty hyperparameters (Figure 4.3). The ensemble forecasts following the DA step are compared to the reference open loop ensemble scenario, which did not implement assimilated data. Results in Figure 4.3 are thus shown in relative terms compared to the open loop forecasts. It is important to recall that the presented results are for short- to medium-term forecasts (up to nine days) whereas some processes are active on much longer periods, such as the snow accumulation and melt processes. In such cases, adjusting state variables might not have an impact on shorter periods. For example, adjusting the snowpack during winter will adjust the amount of snow available in future months, but unless temperatures cross the melting point, they will not have an impact on the resulting short-term hydrological forecasts.

The results show that the DA performs better than the reference open loop ensemble scenario when updating the vadose zone state variable for each of the four seasons when certain combinations of precipitation and temperature uncertainty are applied. It can be seen that lower CRPS values (negative relative values) are attainable mainly by varying the temperature uncertainty within specific ranges during data assimilation. Precipitation uncertainty also plays a role, but it is less sensitive than the temperatures for all seasons during data assimilation for short-term forecasting. A short summary of results is presented here for each state variable.

Vadose zone: The vadose zone state variable is particularly sensitive to the variation in winter temperature. Indeed, a rising temperature will modify the precipitation type and melting when the temperatures are near the melting point, giving more flexibility to find better initial states. In summer, temperatures play an important role due to evapotranspiration rates which lead to a reduction of summer streamflow. Allowing for more uncertainty in the observed temperature values thus allows the model to explore wider variations of simulated flows, leading to better assimilated states. Summer is the season for which the precipitation uncertainty hyperparameter has the largest impact. The interactions between various combinations of precipitation and temperature uncertainty hyperparameters are also the most complex in summer for the vadose zone.

Saturated zone: As for the saturated zone state variable, a good performance of the forecast scenario is seen in winter and fall, with better CRPS values compared to the open loop ensemble scenario for specific combinations of uncertainty hyperparameters. However, for winter, after a certain temperature threshold is attained (here about 6°C), the forecast skill begins to worsen. These findings are not noticeable in summer and spring, where for the saturated zone state variable a variation in temperature leads to no improvement of the ensemble forecasts with a performance similar or slightly better than that of the open loop scenario, even with very large temperature uncertainty of 10°C.

Snowpack: For the snow state variable, the forecast scenarios perform well for small values of temperature in winter, between 2 and 5 °C. After a specific threshold temperature value, the ensemble forecasts diverge until exceeding the CRPS open loop ensemble scenario, probably due to the excessive temperature uncertainty which is no longer realistic, and which forces the model to modify the mass balance too much to minimize the initial error. No clear improvement of the forecast skill is observed in fall for this state variable. The high sensitivity of the ensemble forecasts to temperature in winter and spring for the snow state variable can be explained by the snowpack dynamics, where a small variation in temperature has a direct impact on generating snow melt flows and apportionment of precipitation into rain and snow. This winter influence extends to the saturated zone. In fall, variations in temperature for the snow state variable have little impact on ensemble forecasts because snowmelt flow contributes little to the total streamflow in that season.

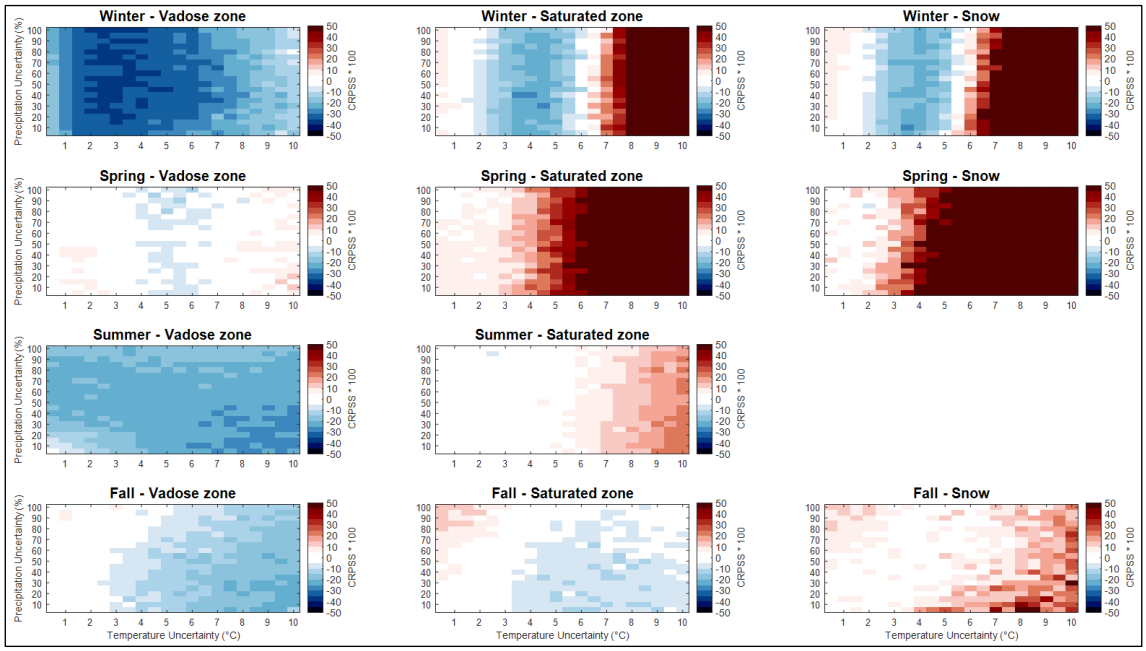


Figure 4.3 CRPSS of the streamflow ensemble forecasts by season when considering variations in temperature from 0.5 °C to 10 °C, precipitation from 5 to 100% and a constant uncertainty value of 5% for streamflow. Values represent relative differences between the assimilated forecasts and the open loop forecasts. Red (blue) values indicate worse (better) performance than the open loop run for the season

The particularly weak sensitivity to precipitation uncertainty in winter can be explained by the fact that the study catchment is snow-dominated with cold climate conditions in winter. In winter, until the beginning of spring, precipitation falls as snow which accumulates over many months prior to generating melting flows at the end of the spring. As spring precipitation is a mix of solid and liquid phases, the forecasts show more variability. In summer and fall, precipitation is in its liquid form and streamflow forecasts are slightly more affected by precipitation uncertainty.

From Figure 4.3, we select a precipitation uncertainty value of 10% for all seasons as it either has no impact beyond that level or is better than larger values in most cases. Furthermore, using larger values unnecessarily generates initial states that could impact longer-term forecasts. Given the threshold values observed for temperature, a temperature value with a seasonal standard deviation of 3, 5, 7 and 8°C is selected for winter, spring, summer and fall, respectively, for further analysis.

### 4.3.2 Influence of the choice of state variables to update in EnKF DA

To further examine the influence of updating state variables on the streamflow forecast skill, all combinations of state variables are evaluated for use in DA for each season using the hyperparameters that were found to be optimal in the previous section.The ensemble forecasts generated with the DA experiments are compared to the ensemble forecasts generated with the individual state variables as in Figure 4.3, as well as in the open loop scenario, as seen in Figure 4.4.

The results reveal that, in winter, the three combinations that include the vadose zone perform well, with CRPS values better than the open loop ensemble scenario. However, choosing to update two or three state variables in winter does not have a clear impact on the accuracy of the ensemble forecasts which are not improved against the DA experiment with one updated

state variable (e.g., vadose zone). In spring, the overall skill of the ensemble forecasts is not improved by adding more state variables. The best case still remains the sole use of the vadose zone as the updated state variable. In summer, the combination of vadose and saturated state variables has similar performance to the single vadose zone results. In the interest of simplicity, the rest of the results are shown using only the vadose zone for summer as well. Similarly to winter, considering a combination of updated state variables in fall does not provide large gains in performance since the ensemble forecasts generated with a single updated state variable are already of relatively good quality. However, using a combination of vadose zone and snowpack state variables does provide a slight increase in forecast skill, thus this combination was preserved for the rest of the study.

Therefore, the results show that the performance of the ensemble forecasts are also related to the choice of the updated state variables for each season and all state variables should not be systematically updated for every season. According to these results, the state variables to update in order to get the best performance of the ensemble forecasts with a low CRPS are:

1. Winter: either the vadose zone on its own, or any combination that includes the vadose zone. But for simplicity, the vadose zone taken individually is recommended;
2. Spring: the vadose zone on its own or in combination with the saturated zone;
3. Summer: the vadose zone on its own or in combination with the saturated zone;
4. Fall: any combination that includes the vadose zone, with a small preference for the combination of the vadose zone and the snowpack.

To keep things as simple as possible, a user could preserve almost all the skill by focusing only on the vadose zone during the updating of state variables. These best sets of state variables to update in DA are considered for further analysis in the next section.

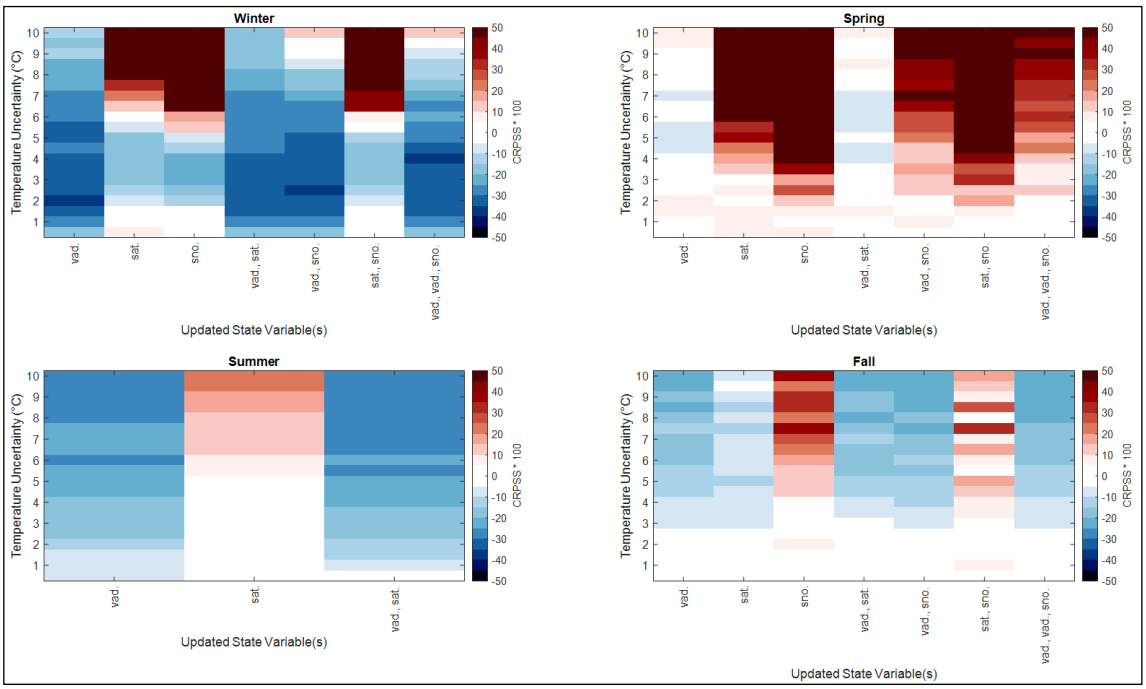


Figure 4.4 CRPSS score of the streamflow ensemble forecasts by season when considering different combinations of state variables with constant seasonal values of the precipitation (10%) and streamflow (5%) with varying temperature uncertainty hyperparameters. [Abbreviations in this figure = vad.: vadose zone; sat.: saturated zone; sno.: snowpack]

### 4.3.3 Insights on the best hyperparameter and state variable combinations per season

Using the best hyperparameter values and the best combinations of updating state variables applied for each season, we are able to provide a DA recipe (Table 4.1) in order to gain as much forecast skill as possible.

Figure 4.5 and Figure 4.6 present the skill of ensemble forecasts generated after application of DA using the recommended recipe at the forecast lead-time of day 1 and day 9, respectively. Boxplots contain the results of all forecast days for each season (see Table 4.1). The lower the

CRPS value, the better the forecast skill for the given lead-time. As expected, the ensemble forecasts are clearly improved with the recommended recipe at the lead time of day 1, particularly in winter, summer, and fall, with a low ensemble spread and uncertainty compared to the open loop ensemble scenario and for which the results are statistically significant at the 5% level using a Wilcoxon non-parametric test. For spring, the CRPS variability is lower overall with the application of the DA recipe, but the median is similar to the open loop run. It is also clear that for some seasons, such as spring and summer, the forecasting skill quickly falls to open loop levels or worse for the longer lead times, as seen in Figure 4.6. In Figure 4.6, at the 9$^{th}$ day of lead-time, only the winter season shows significant differences between both scenarios. This shows the benefits of using a DA recipe for each season in order to prolong the performance of the ensemble forecasts for larger lead times for the more responsive seasons. The improvement over the open loop ensemble scenario can be quantified using the CRPSS metric for each day of lead-time. This is presented in Figure 4.7, with CRPSS values for each lead-time and for each season.

Finally, by pooling all forecast events together, it is possible to investigate the added value of the DA scheme for streamflow forecasts for each of the lead times from 1 to 9 days, but over the entire multi-year period. In this case, the DA recipe is followed by changing the hyperparameters for each new season according to the recommendations in Table 4.1. These results are shown in Figure 4.8.

It can be seen that the median and lower quantile forecast CRPS values are better than those using an open loop approach for all lead-times. Much of these gains are due to the winter period, which responds very well to the DA. However, even the 75$^{th}$ quantile seems to be slightly better using DA over the entire year, although starting at day 6, these results start to be less significant or indifferent.

Table 4.1 Summary of the recommended DA recipe for each season

| | | Winter | Spring | Summer | Fall |
|---|---|---|---|---|---|
| Hyperparameter | Temperature uncertainty | 3°C | 5°C | 7°C | 8°C |
| | Precipitation uncertainty | 10% | 10% | 10% | 10% |
| | Streamflow uncertainty | 5% | 5% | 5% | 5% |
| Combination of updated state variables | | vadose zone | vadose zone | vadose zone | vadose zone snowpack |
| Number of forecast days in the period | | 191 | 101 | 122 | 122 |

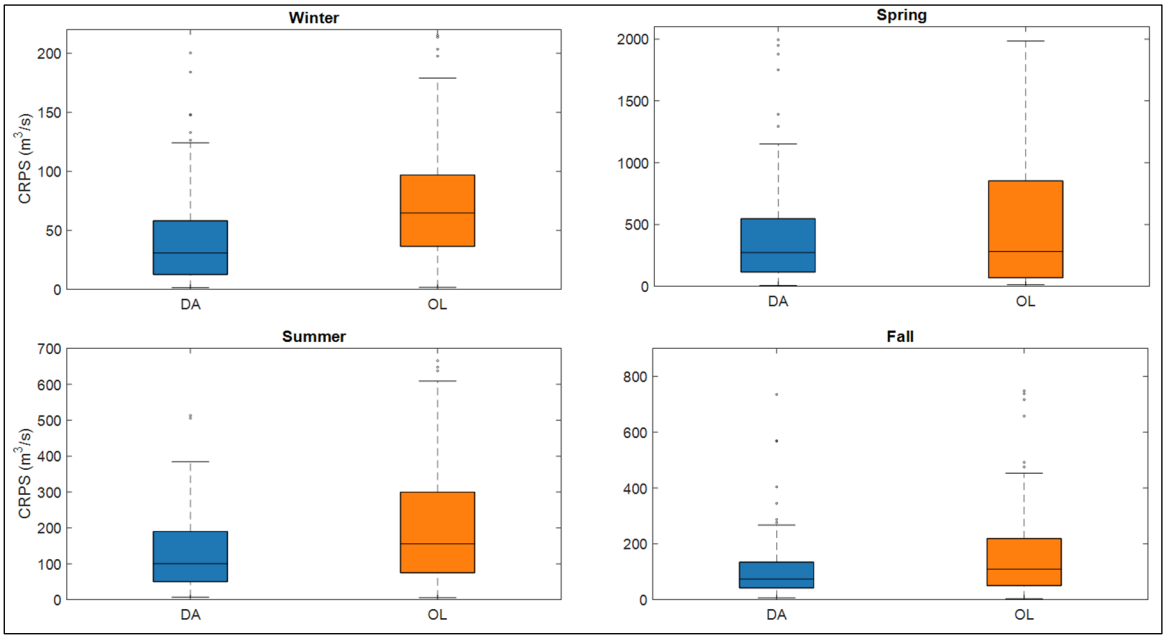


Figure 4.5 CRPS values of the streamflow ensemble forecasts for data assimilation (DA; blue) and open loop ensemble (OL; orange) by season at day 1 when

considering the best combination of state variables and the optimal hyper-parameter values by season

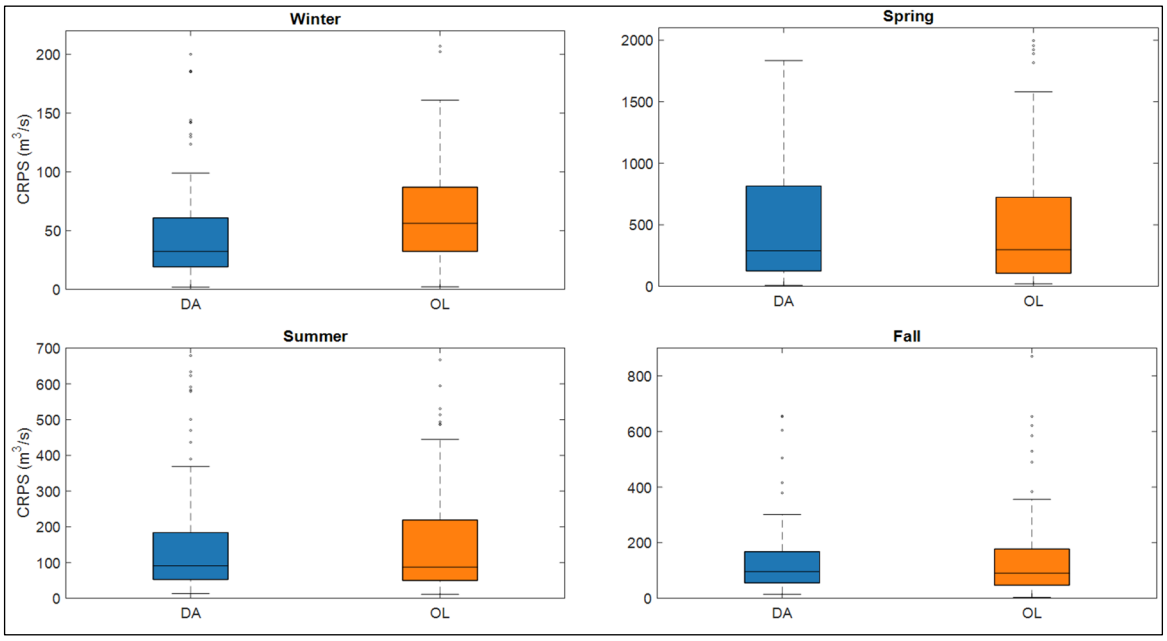


Figure 4.6 CRPS values of the ensemble for data assimilation (DA; blue) and open loop ensemble (OL; orange) streamflow forecasts by season at day 9 when considering the best combination of state variables and the optimal hyper-parameter values by season.

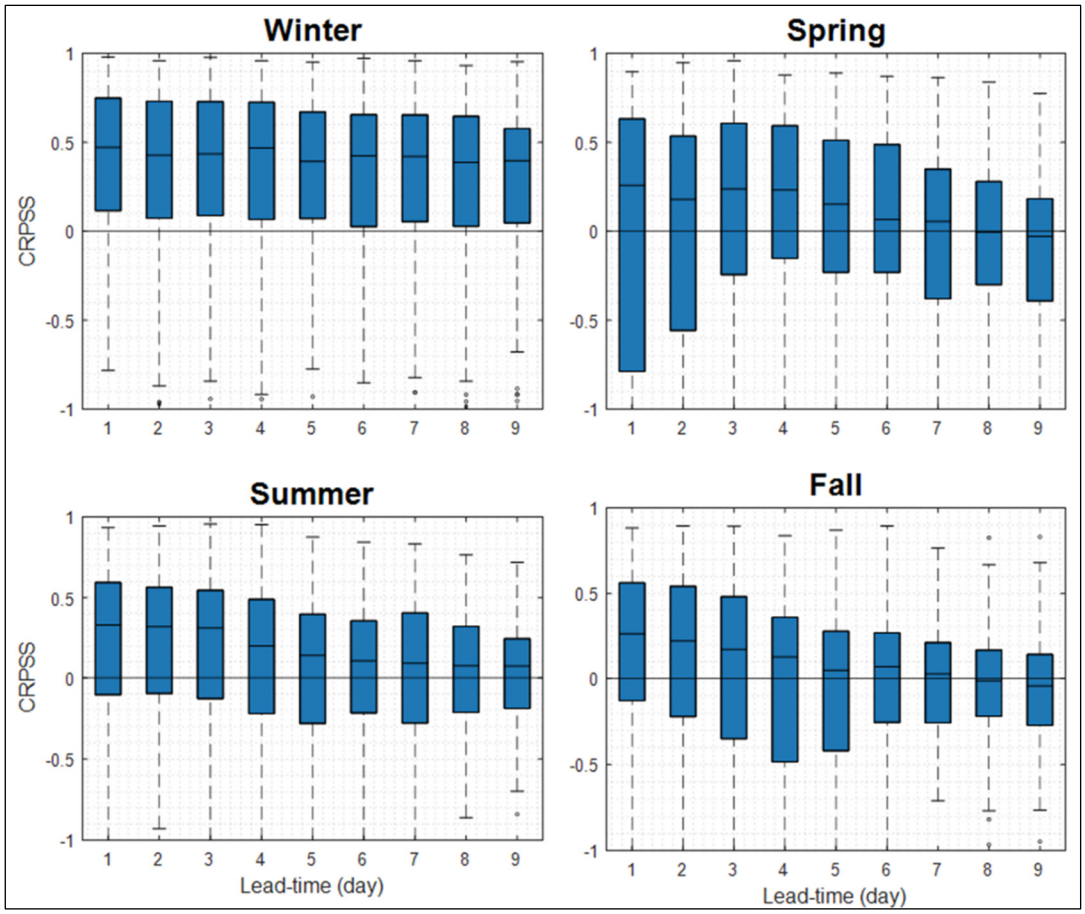


Figure 4.7 CRPSS results. An overview of seasonal comparative performance of forecasts using DA and open loop initial states. Positive values indicate better forecasts than the open loop forecast

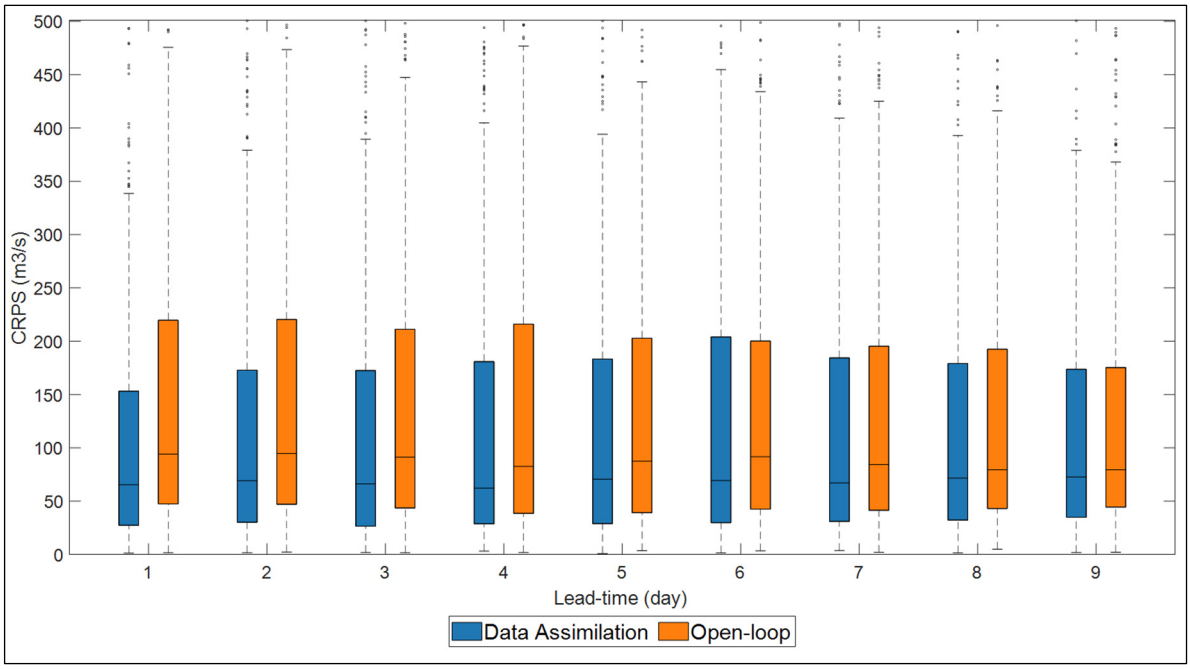


Figure 4.8 CRPS values of the annual ensemble streamflow forecasts during the entire lead-time when considering the best combination of state variables and the optimal hyper-parameter values for each season

### 4.3.4 Forecast performance evaluation

To assess the performance of the EnKF DA method, the ensemble streamflow forecasts generated with the DA scheme and the open loop approach are compared to the observed streamflow for a 9-day forecast (Figure 4.9), which is the maximum lead-time available in our weather forecast data. The results show the general added value of the DA recipe to the overall quality of ensemble forecasts in winter for lead-times up to 9 days, which surpasses that of the open loop ensemble. In spring, this is true for shorter lead-times (below 5 days) after which the performance of the ensemble forecasts decreases with increasing lead-time to get closer to that of the open loop ensemble. The summer DA scheme increases the spread of the ensemble forecasts for shorter lead-times (below 6 or 7 days) which allow it to get closer to the observed streamflow compared to the open loop ensemble. The comparison of DA-ensemble forecasts with the open loop ensemble in fall shows that the DA generates skillful forecasts for lead-

times between days 2 to 9 days depending on the day and hydrometeorological conditions. However, as seen in Figure 4.7, after about 5 days the effects of DA become marginal, consistent with the findings of (Samuel et al., 2019).

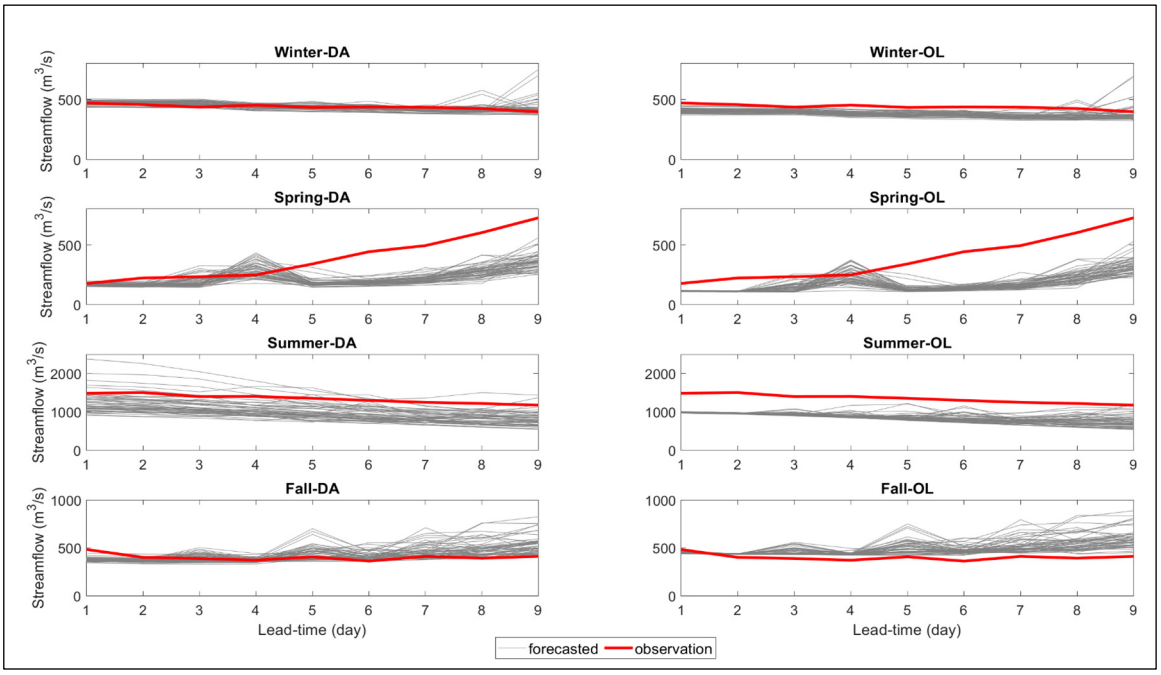


Figure 4.9 Examples of forecasts before and after application of DA for each season, including observed (red) and forecasted (gray) streamflow ensembles during over the 9-day lead-time when considering the best seasonal DA recipe (hyperparameter set and state variables)

## 4.4 Discussion

This work has provided insights regarding the seasonal forecasting accuracy of streamflow when the EnKF-based assimilation of input (precipitation and temperature) and output (streamflow observations) is performed within the operational CEQUEAU hydrological model over the Lac-Saint-Jean catchment (Canada). The focus is placed on the estimation of forecast improvement for multiple lead-times, and how the different EnKF hyperparameters and state variables need to be seasonally perturbed to construct a more accurate forecasting platform.

In general, the DA scheme has a positive impact on the streamflow forecasts for lead-times shorter than 6 days. The effect of DA decreases with longer lead-times, except in winter where Figure 4.3 the benefit of the DA extends until the end of the forecast window at day 9 (Figure 4.8). The ensemble forecasts after the 6-day lead-time depict little difference compared to the reference open loop scenario, which indicates that the DA approach has little effect when the lead-time is longer than the concentration time of the catchment, which is of about 5 days on the Lac-Saint-Jean catchment depending on the season and hydrological conditions. Chen et al. (2013) indicate that the DA leads to no improvement on streamflow forecasting when the lead-time exceeds the flow routing time of the catchment, which represents a physical limit of the forecasting process. Patil and Ramsankaran (2017) show that the persistence of streamflow forecast improvements is longer when the subsurface flow component is updated for their model. Our results corroborate these findings as the vadose zone state variable is the most sensitive and impactful in all our tests and is deemed the most important for all seasons.

The results show the general added value of a season-conditioned DA approach to the overall quality of ensemble forecasts. When evaluating the quality of ensemble forecasts with the CRPS and CRPSS values, the benefits in terms of skill depend first on the hyperparameter sets. Accuracy of ensemble forecasts is achieved by using important perturbations of the temperature hyperparameter, with up to 9 °C uncertainty on the observations. Precipitation and streamflow play very limited roles in this regard, which is somewhat surprising given that precipitation is typically more difficult to measure than temperature. Nonetheless, larger values of precipitation uncertainty are detrimental to forecast skill in summer. This illustrates the level of uncertainty associated with the meteorological input data, but could also be a means to compensate for other biases and flaws present in the rest of the hydrological model and forecasting chain. A fundamental limitation is the representativeness of such observation perturbations which can be somewhat unrealistic. However, a hyperparameter set with low input temperature perturbations does not provide ensemble forecasts that are as skillful (Figure 4.3). In most cases, the efficiency of the DA scheme is seen right after the model spins up and

the duration of efficiency for longer lead-times varies between seasons. This shows that other predictive uncertainties need to be considered to obtain a reliable forecasting system. Some studies address the issue of the estimation of the ensemble size (not explored here), as the efficiency of the DA procedure through the tuning of the hyperparameters is linked to the ensemble size (Moradkhani et al., 2005; Thiboult & Anctil, 2015). The authors show that an ensemble size of at least 40 members allows sampling error to be limited. Most hydrological forecasting studies rely on a 50-member ensemble as a compromise between stochastic errors and computational cost during the sampling of the distributions of the state variables (Valdez et al., 2022). Here, a 50-member ensemble is considered for seasonal streamflow forecasting over the Lac-Saint-Jean catchment.

We also show that the selection of state variables to update plays a role in the forecast skill improvement. As the chosen state variables (vadose zone, saturated zone, and snowpack) are representative of reservoirs located at different depths in the CEQUEAU model, who interacts with them differently according to the seasons, the combined or individual updating of these state variables can influence the seasonal ensemble forecasts of streamflow in a different manner. As shown in (McMillan et al., 2013) and (Rakovec et al., 2015), the issue of the number of state variables to update is challenging since updating all model state variables does not systematically imply the best findings in terms of reliability of streamflow forecasts. It is also shown that applying DA techniques such as EnKF could improve short-term forecast skill at the cost of long-term water balance disruptions, which could lead to much worse long-term forecasts (Mai et al., 2020). This is in agreement with our results, in which we can see that the spring CRPS values become worse than that of the open loop over a span of 9 days (Figure 4.5 and Figure 4.6). In that particular case, it is likely that the DA modifies the states in such a way that the short-term error is minimized but introduces new errors in the mass balance that affect longer-term forecasts.

## 4.5 Conclusion

In this study, an EnKF-based assimilation of inputs (temperature and precipitation) and output (streamflow) is applied to the operational CEQUEAU hydrological model to improve short-term streamflow forecasting performance over the Lac-Saint-Jean catchment (Canada). In order to enhance the efficiency of the assimilation, a new ensemble is created by perturbing the EnKF hyperparameters and by selecting the most relevant model state variables to update for each of the four seasons. This work compares the 50-member ensemble of streamflow simulations obtained through perturbed values of the EnKF hyperparameters and state variables with the reference open loop scenario without assimilation. A sensitivity analysis is conducted to evaluate the influence of the hyperparameters and model state variables on the reliability of the seasonal forecasting performance.

The results indicate that the use of a seasonal DA scheme with fixed hyperparameters (P: precipitation, T: temperature, Q: streamflow) and best state variables combination (vadose zone and/or saturated zone and/or snowpack) by season leads to the overall improvement of ensemble streamflow forecasts. Ensemble forecast skill decreases with longer lead-times as the effects of the DA wear off after the first six days. Despite this, the gain provided by EnKF over the open loop scenario is substantial, especially in winter. These results imply that conducting a detailed testing of all possible combinations to identify the best seasonal performing EnKF implementation could help operational forecasters better target the assimilation hyperparameters for their specific need. For future forecasting studies, we encourage performing a detailed sensitivity analysis which addresses the issues of the seasonal selection of the EnKF hyperparameter and state variable selection of the forecasting system to ensure the EnKF relevance.

Overall, the findings of this study show the strong seasonal sensitivity of the EnKF hyperparameters and the model state variables. A judicious seasonal implementation of the

EnKF assimilation scheme has significant potential to improve streamflow forecasts for operational forecasting centers and water resources system managers.

## 4.6 Acknowledgements

This study was partially funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) under the Collaborative Research and Development grant CRDPJ-522126-17. The authors would like to thank Rio Tinto for sharing their hydrometeorological data on the Lac-Saint-Jean catchment. The authors are also grateful to ECMWF for providing access to the historical forecast data from their MARS computing and archiving facilities. The authors are also grateful to the anonymous reviewers whose suggestions and comments helped improve the paper and shape it into its current form.

# CHAPTER 5

# COMPARING A LONG SHORT-TERM MEMORY (LSTM) NEURAL NETWORK WITH A PHYSICALLY-BASED HYDROLOGICAL MODEL FOR STREAMFLOW FORECASTING OVER A CANADIAN CATCHMENT

Behmard Sabzipour [a], Richard Arsenault [a], Magali Troin [a, b], Jean-Luc Martel [a], François Brissette [a], Frédéric Brunet [a], Juliane Mai [c,d,e]

[a] Hydrology, Climate and Climate Change Laboratory, École de technologie supérieure, Montréal, Canada.
[b] HydroClimat| TVT, Maison du Numérique et de l'Innovation, Toulon, France
[c] Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, Canada.
[d] Department of Computational Hydrosystems, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany
[e] Center for Scalable Data Analytics and Artificial Intelligence - ScaDS.AI, Humboldtstraße 25, 04105 Leipzig, Germany

Paper submitted to *the Journal of Hydrology*, February 2023

## Abstract

Streamflow forecasting is crucial in water planning and management. Physically-based hydrological models have been used for a long time in these fields, but improving forecast quality is still an active area of research. Recently, some artificial neural networks have been found to be effective in simulating and predicting short-term streamflow. In this study, we examine the reliability of Long Short-Term Memory (LSTM) deep learning model in predicting streamflow for lead-times of up to ten days over a Canadian catchment. The performance of the LSTM model is compared to that of a process-based semi-distributed hydrological model, with both models using the same weather ensemble forecasts. Furthermore, the LSTM's ability to integrate observed streamflow on the forecast issue date is compared to the DA process required for the hydrological model to reduce initial state biases. Results indicate that the LSTM model forecasted streamflows are more reliable and accurate

for lead-times up to 7 and 9 days, respectively. Additionally, it is shown that the LSTM model using recent observed flows as a predictor can forecast flows with little error in the first forecasting days without requiring an explicit DA step, resulting in improved skill compared to both the open-loop and assimilated distributed hydrological model.

**Keywords:** Long short-term memory (LSTM); hydrological forecasting; data assimilation; ensemble forecasting; deep learning.

## 5.1 Introduction

Accurate, reliable, and easily understandable hydrological forecasts are crucial for a wide range of users in water- dependent sectors, such as agriculture, hydropower, and floodplain management (Anghileri et al., 2016; Boucher et al., 2012; Cassagnole et al., 2021). As a result, forecasting streamflow has been the focus of numerous studies since the mid-1970s (Day, 1985; Twedt et al., 1977), and has seen an increasing amount of attention in recent decades (Troin et al., 2021) as demands for water resource management and natural disaster mitigation have risen substantially.

There are two main methods used to forecast streamflow: the first is the use of dynamical (or process-driven) hydrological models, which range from conceptual and lumped to physically-based and distributed models; the second is the use of data-driven statistical models such as machine learning (ML), artificial neural networks (ANNs) and autoregressive models (Rajagopalan et al., 2010). These approaches can provide an ESP system when ensemble weather forecasts are used as inputs. However, dynamical hydrological models are often limited by the availability of data required for their implementation, such as soil type and depth; or by their simplistic process representations (Damavandi et al., 2019). These issues can be overcome using deep learning approaches, which can lead to reliable simulations of hydrologic systems even when the underlying physical processes are not explicitly taken into account (Maier et al., 2010). A particular type of ANNs that has become increasingly popular in the

last years in the field of hydrology, is long short-term memory (LSTM) network, due to its ability to process sequential data and time series (Shen & Lawson, 2021; J. Zhang et al., 2018).

LSTMs, introduced by Hochreiter and Schmidhuber (1997), are an advanced type of recursive neural networks (RNNs) designed to learn sequence (temporal) data and their long-term dependencies outperforming the latter (Kratzert et al., 2018). An LSTM consists of a memory cell, which is a neuron with a self-recurrent connection, and three nonlinear gates that control the movement of information within and outside the cell. The forget gate, or cell memory, preserves information for a long time, allowing the LSTM to learn long-term dependencies that other RNNs cannot. As a result, LSTMs can overcome the issues of short-term memory and vanishing gradients that traditional RNNs face (Xu et al., 2020).

The potential use and benefits of LSTMs in the field of hydrology have recently begun to be explored (Arsenault et al., 2023; Hu et al., 2018; Hunt et al., 2022; Khoshkalam et al., 2023; Kratzert et al., 2018; Sahoo et al., 2019; J. Zhang et al., 2018). For example,J. Zhang et al. (2018) evaluated the performance of various RNNs architectures in simulating water levels in Norway and found that the LSTM is better suited for multi-step-ahead forecasts than other architectures without cell memory. Kratzert et al. (2018) compared the performance of the LSTM and the SAC-SMA hydrological model (Sorooshian et al., 1993) for simulating long-term streamflow over 241 catchments in North America and found that the former outperformed the latter, highlighting the potential of LSTMs as regional hydrological models. Kratzert et al. (2018) and Lees et al. (2021) also noted the possibility of applying LSTMs in regions other than the one used for training. Hunt et al. (2022) evaluated the performance of the LSTM in predicting streamflow in various climate regions of the United States and compared it to the Copernicus Emergency Management Service (CEMS) physics-based GloFAS. The authors reported that LSTM generated skillful forecasts that outperformed the raw and bias-corrected GloFAS forecasts up to a 5-day lead-time. In this context, Hu et al. (2018) attributed the better performance of LSTM compared to conceptual and physical-based hydrological models to the feature of the forget gate. However, other studies have reported the

difficulty using LSTMs for predicting streamflow during extreme conditions, as well as for catchments with complex groundwater-river interactions and human abstractions (Kratzert et al., 2018; Lees et al., 2021). Arsenault et al. (2023) showed that LSTMs systematically outperformed traditional hydrological models in a regionalization experiment performed over 148 North American catchments. Nevo et al. (2022) developed a flood forecasting model using LSTMs and found that it was significantly better than more traditional multiple linear regression models in forecasting river water stages in India and Bangladesh.

Recently, some studies have begun investigating the use of encoder-decoder (ED) LSTMs, also known as sequence-to-sequence or Autoencoder (AE) LSTMs. Kao et al. (2020) found LSTM-EDs are reliable in converting rainfall sequences to runoff sequences and suitable for forecasting hourly floods. Zhang et al. (2022) showed that LSTM-EDs perform better than a conceptual model for predicting floods in ungauged catchments. These consist of two RNNs, an encoder and a decoder. The encoder converts a variable-length sequence to a fixed-length vector , and the decoder converts the fixed-length vector back to a variable-length sequence (Ghimire et al., 2022). There is a hidden layer between these two steps that processes the encoded data (Cho et al., 2014; Wang et al., 2016), allowing the sequences of observed data to be used directly to predict multiple timesteps at a time and using previous states for a given forecast. AEs have recently been applied to streamflow forecasting, by(Kao et al., 2021) using an LSTM autoencoder for flood inundation forecasts to generate multistep-ahead regional inundation maps; Ponnoprat (2021) using a seasonally integrated autoencoder combined with a LSTM for predicting short term dynamics and seasonality of daily precipitation; and (Girihagama et al., 2022) using an attention-based Encoder-Decoder LSTM to improve forecasting skill over ten catchments in Canada. These methods should also be considered in future research on LSTM-based forecasting (Provotar et al., 2019), although they remain complex and more difficult to implement than simpler LSTM models for forecasting.

Incorporating and integrating observational data is important for better performance of hydrological forecasting, as it helps adjust model states such that the model represents actual

hydrological conditions as best as possible. This is done by applying methods such as data assimilation (DA) and regressions for state updating (Brajard et al., 2020; Fang & Shen, 2020; Nearing et al., 2022). Nearing et al. (2022) discussed ways of integrating data in LSTMs to represent missing streamflow data for gauged and ungauged basins. They used a regression method and variational DA method. The integration part was added to the cell state of LSTM, i.e., the one which is the recursive state of LSTM, in order to over-train observed streamflow data which are spared. They found DA has advantages over autoregression, since it is able to deal with up to 50% of missing data. They reported DA worked better when used for catchments including both gauged and ungauged basins. (Brajard et al., 2020) reported successful results for combining the EnKF with a surrogate model of a neural network. Fang and Shen (2020) used LSTMs to forecast soil moisture using satellite data using a Data Integration kernel to assimilate and update models with the most recent available observations showing that it was effective to reflect unseen processes in inputs, such as floods.

This study aims to evaluate the potential of LSTM to simulate and forecast daily streamflow over the Lac-Saint-Jean (LSJ) catchment, located in the province of Quebec, Canada. In this region, a large portion of the streamflow comes from snowmelt, making it an ideal environment for testing LSTM forecasts in snowmelt-dominated catchments. As there are only a few applications of LSTM forecasting in Nordic regions, the architecture and reliability of this deep learning technique needs to be investigated to identify the benefits for streamflow forecasting as an operational deployment in this region. The strong snowpack dynamics and the importance of capturing such long-term hydrological processes make it more challenging than in regions with more uniform weather. (Girihagama et al., 2022) used an ED LSTM with an attention mechanism to provide streamflow forecasts in ten river catchments in the Great Lakes region in Canada. Their study concluded that this variant of LSTM was able to provide excellent forecasting results for up to five days of lead-time. (Khoshkalam et al., 2023) employed transfer learning with LSTM, leveraging meteolorogical and physiographic data, to enhance forecast skill in snow-dominated regions. By training the model on data-rich regions and

emphasizing the inclusion of recent streamflow data, the study underscored the significance of integrating multiple data sources for improved predictions.

In particular, this study seeks to address the following research questions:

- How well do LSTM-based models forecast streamflow over a catchment where the hydrologic response is dominated by snowmelt?
- How does LSTM-based model performance compare with the operational forecasting system (a semi-distributed hydrological model combined with a DA scheme) used as a benchmark?

The experimental design, including the study area and data used to train and evaluate the models, is described in section 5.2. The method, including a description of the LSTM and the semi-distributed hydrological model, is presented in Section 5.3. The results are analyzed in Section 5.4 followed by a discussion in Section 5.5. Concluding remarks appear in Section 5.5.2.

## 5.2 Experimental design

### 5.2.1 Study area

This study was conducted on the LSJ catchment in the province of Quebec, Canada (Figure 5.1). The catchment has an area of 45000 $km^2$ and is used for hydropower generation by the Rio Tinto corporation for aluminum smelting. The catchment is made up of nine monitored sub-catchments, which drain into the reservoir, as shown in Figure 5.1a. The figure also shows the catchment location (Figure 5.1d). The sub-catchment “other tributaries” is made up of a series of small rivers and streams that all flow to the reservoir but are ungauged. The annual average precipitation on the LSJ catchment is 1000 mm, of which 34% falls as snow, as measured by Rio Tinto’s weather monitoring network stations. This provides substantial inflows to the Lac-St-Jean reservoir, a 1000 $km^2$ reservoir that can store up to 4550 $hm^3$ of

water for the hydropower generating station (Arsenault & Côté, 2019). Spring peak flows are associated with snowmelt events, while high flows in summer and fall are related to precipitation events. Mean annual streamflow is about 900 $m^3s^{-1}$ (Bergeron et al., 2021).

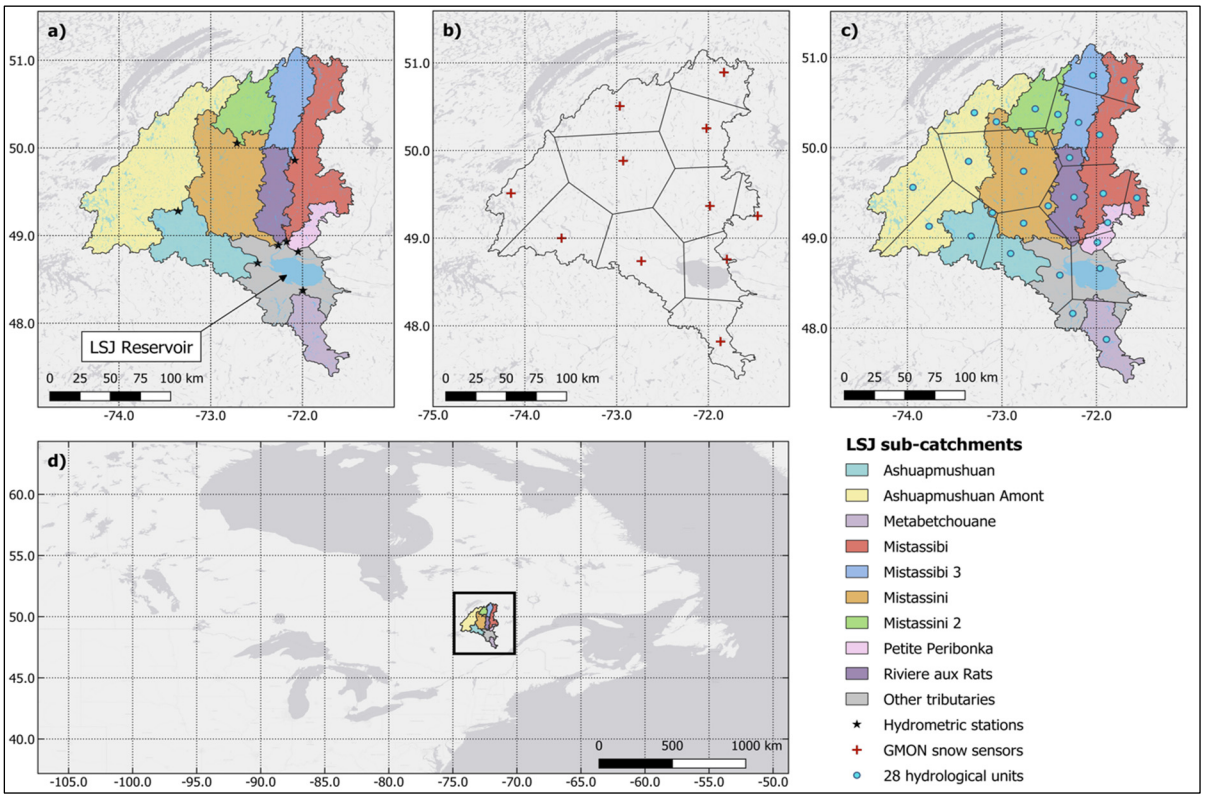


Figure 5.1 LSJ catchment and its ten contributing sub-catchments, including the "other tributaries" which are the sum of all smaller ungauged rivers flowing into the LSJ reservoir (a). Hydrometric stations are represented by stars (a). The eleven gamma ray monitors (GMON) are represented by red crosses over the Thiessen polygons they generate (b). The intersection of the ten sub-catchments and the eleven Thiessen polygons represent the 28 hydrological response units (HRU) of the CEQUEAU semi-distributed hydrological model used in this study, represented by a blue circle (c). The location of the LSJ catchment in Quebec, Canada, is shown within Eastern North America (d)

### 5.2.2 Datasets

In this study, multiple datasets were used, including streamflow data as well as observed and forecasted weather data. This section describes all datasets and their pre-processed methodology.

#### 5.2.2.1 Observed hydrometeorological data (used for calibration of hydrologic model)

The observed hydrometeorological data were provided by Rio Tinto, the operator of the hydropower generating station and owner of water rights for the LSJ system. The data covered the period from January 1954 to December 2019 and included daily minimum and maximum temperatures and precipitation from a network containing 16 weather stations distributed throughout the catchment.

In addition to the weather data, Rio Tinto also provided inflows to the main reservoir, which were derived through mass balance calculations by evaluating changes in the reservoir level at various locations along with known outflows from turbines and spillways. This was necessary as multiple rivers and tributaries flow into the reservoir, making a significant portion of the flows ungauged. However, the mass-balance derived inflows can be noisy at times due to wind displacing the water surface, which can bias storage volumes over short periods (Loiselle et al., 2021). To address this, a three-day moving average was applied to smooth out variations in inflows, which did not change the total water balance over longer horizons, or the timing of events in a significant manner. The inflows to the reservoir are the target of the forecasting procedure in this study.

#### 5.2.2.2 Ensemble weather forecast (used for forecasting with hydrologic and LSTM model)

To assess the performance of the semi-distributed hydrological model and the LSTM model in forecasting mode, the 50-member operational ensemble weather forecast data was obtained from ECMWF IFS using the Meteorological Archival and Retrieval System (MARS) archive (https://www.ecmwf.int/en/forecasts/dataset/operational-archive; Grawinkel et al., 2015). The variables of total precipitation, maximum and minimum temperature were available at a spatial resolution of 0.2° x 0.2° on a 6-hourly time step twice per day (0Z and 12Z). For this study, only the 0Z forecasts were utilized and aggregated to a daily time step. The forecast lead-times from one to ten days were downloaded, but the last day was truncated due to the five-to-six-hour time zone offset that made the tenth day unavailable for the entire period for the study location, resulting in an effective nine-day ensemble weather forecast. The forecasts were downloaded for the period of January 2015 to December 2019. Data prior to 2015 were not included due to a major update of the ECMWF integrated forecasting system in 2015, which caused a change in weather forecast statistics that would not be representative of the more recent model versions. The ensemble forecasts of precipitation and temperature were then spatially aggregated to the scale of the 28 distributed hydrological model sub-regions for the hydrological model and over the entire catchment as input to the LSTM model. Note that the raw forecasts were used, and no bias-correction was implemented at this stage.

#### 5.2.2.3 Reanalysis data (used for training of LSTM model)

In this study, a second set of pseudo-observed meteorological data, known as reanalysis data, was used to train the LSTM model. This data is from the ECMWF fifth generation reanalysis (ERA5; Hersbach et al. (2020)). The reanalysis data used to provide observations on the historical period, so that the LSTM model could be used to forecast data with the same variables in the forecasting period. The station-based observations provided by Rio Tinto (Section 5.2.2.1) did not contain the desired spatial and temporal coverage of wind, solar

radiation, and pressure variables for this training, thus ERA5 data was used instead. The variables used from ERA5 are the same as for the forecast data, but at an hourly time step on a 0.25° x 0.25° resolution. The ERA5 data was aggregated at the daily scale and was spatially aggregated over the entire catchment, to maintain consistency with the spatial and temporal scales of the ensemble weather forecast data. The data is available from January 1979 to the present day with a latency of approximately five days, which was deemed sufficient for training the LSTM model in this study.

## 5.3 Methods

This study aims to evaluate the ability of the LSTM neural network model to forecast streamflows on a large, snowmelt-dominated catchment. A traditional semi-distributed hydrological model is also used as a comparison over the same catchment and time periods. The methods used to achieve the study objectives are described in detail in this section, including the forecast evaluation metrics (Section 5.3.1), the traditional hydrological modeling and forecasting (Section 5.3.2), and the LSTM model training and forecast testing (Section 5.3.3).

### 5.3.1 Performance evaluation criteria

The performance of the different streamflow forecasts is evaluated using the Kling-Gupta Efficiency (KGE; Gupta et al. (2009), Continuous Ranked Probability Score (CRPS;(Hersbach, 2000) and Mean Absolute Error (MAE; (Mather & Johnson, 2016) at different lead-times (from days one to nine) during the forecast.

The KGE, which is unit less, is defined as:

$$KGE = 1 - \sqrt[2]{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2} \tag{5.1}$$

where r is the Pearson correlation coefficient; β is the bias ratio ($\mu_{q_{sim}}/\mu_{q_{obs}}$); and γ is the variability ratio ($\sigma_{q_{sim}}/\sigma_{q_{obs}}$) where $\mu_{q_{sim}}$ and $\sigma_{q_{sim}}$ are the mean and standard deviation of the forecasted (simulated) streamflow and $\mu_{q_{obs}}$ and $\sigma_{q_{obs}}$ are the equivalent for the observed streamflow. A perfect forecast would have a KGE of 1, while suboptimal forecasts show KGE values lower than 1.

The CRPS is a suitable metric for probabilistic forecasts. It measures the average distance between the observed probability density function ($F(q_{obs})$), and the ensemble forecast probability density function ($F(q_{sim})$), given by:

$$CRPS = \frac{1}{T}\sum_{t=1}^{T}\int_{-\infty}^{+\infty}\left[F(q_{sim_t}) - F(q_{obs_t})\right]^2 \, dq \tag{5.2}$$

where $T$ is the total amount of time steps. The CRPS ranges from 0 to + ∞, with zero being a perfect forecast. For streamflow, the units of CRPS are $m^3s^{-1}$. CRPS has the same dimension as $q$ ($dq$).

Finally, CRPSS is used to compare two forecasts:

$$CRPSS = 1\text{- }(CRPS_{forecast} / CRPS_{reference}) \tag{5.3}$$

where $CRPS_{forecast}$ is the CRPS of the new tested forecast and $CRPS_{reference}$ is the CRPS of a reference forecast. A CRPSS (unitless) of 1 indicates that the forecast has perfect skill, a value of 0 indicates that the forecast has no benefit over the reference, and a negative value indicates that the forecast is less accurate than the reference.

MAE measures the difference between the observed and forecasted results but in a deterministic (rather than ensemble) setting. The MAE is the equivalent of the CRPS when a single member is used, and therefore the CRPS can be seen as an extension of MAE to ensembles. MAE is an average of the absolute error, as:

$$MAE_{(t)} = \frac{\sum_{i=1}^{N} |q_{sim_i}(t) - q_{obs}(t)|}{N} \quad (5.4)$$

where $q_{obs}$ is the observed streamflow, $q_{sim_i}$ is the simulated streamflow (which also means pseudo-forecasted as well) for the $i$-th member of the ensemble of size N at lead-time $t$. The optimal value for MAE is 0.

### 5.3.2 The CEQUEAU hydrological model

The CEQUEAU hydrological model is used as a benchmark for evaluating model performance statistics in this study. CEQUEAU is a semi-distributed model that incorporates two conceptual reservoirs to simulate key hydrologic processes: a surface snowpack reservoir to simulate snow accumulation and melt processes, an unsaturated zone reservoir that describes the soil infiltration and interflow processes, and groundwater storage and base flows (Morin & Paquet, 2007). Evapotranspiration is estimated using the Oudin method (Oudin et al., 2005) and snow processes are modeled with the degree-day CEMANEIGE model (Valéry, 2010).
The schematic for the CEQUEAU model is provided in Appendix I.

The catchment is divided into 28 hydrologic response units, which correspond to the intersection of the influence area of GMON snow monitoring stations and the 10 hydrologic sub-catchments (Figure 5.1a-c). The simulation of vertical fluxes is independently performed on each of the 28 sub-regions, as described in Mai et al. (2020). Unit hydrographs are then employed to route flows to the downstream sub-catchments until they reach the outlet. CEQUEAU only requires daily precipitation and average daily temperature as meteorological inputs. The model was manually adjusted and calibrated by Rio Tinto, which provided their operational streamflow forecasting model for this study. The model was set up using data from January 1954 to December 2014 and has been their primary hydrological forecasting tool since.

CEQUEAU was then employed to generate streamflow forecasts for the period of January 2015 to December 2019, both in open-loop mode (i.e., without DA) and using an implementation of EnKF Evensen (1994) DA procedure. The EnKF used in this study was optimized for forecasting performance on the LSJ catchment and was implemented according to the procedure described in CHAPTER 4, which is summarized in the supplementary material (Section 5.8.1). CEQUEAU was forced with observed meteorological variables (Section 5.2.2.1) until the forecast date using the EnKF to maintain robust state variables up to the day the forecast was issued. At this point, the hydrological model was run using the ECMWF ensemble weather forecasts (Section 5.2.2.2) as meteorological input to generate the ensemble streamflow forecasts. DA was performed every three days to increase computation efficiency and to prevent the model's initial states from drifting too far away from the observations. Streamflow forecasts generated by CEQUEAU (with and without DA) were only generated on the dates that DA was performed to ensure the best possible initial states while allowing a small temporal gap between streamflow forecasts to prevent strong autocorrelations between successive forecast dates.

### 5.3.3 LSTM network

In this study, a single-layer LSTM model with a variable number of units was chosen as the model structure. The setup details and hyperparameters of the LSTM are summarized in Table 5.1. Different hyperparameters and model structures were considered for each lead-time during the model training and were tested to obtain the best results as measured by the KGE metric. The dropout rate, number of epochs, batch size, and number of LSTM units hyperparameters were modified and optimized for each lead-time through trial-and-error to improve performance on the validation period. The dropout rate was included to randomly turn off a certain percentage of LSTM units during training, in order to reduce overfitting and improve modeling performance during forecasting. The number of epochs, or the number of times the model sees the full dataset during the training, and batch size, or the size of the dataset sub-

sample used to estimate the gradient descent, were adjusted to speed up the training and ensure its convergence. This led to a larger number of epochs for longer lead-times.

The LSTM models were trained and tested using multiple meteorological variables from the ERA5 reanalysis dataset, including precipitation, temperature, wind, surface pressure, net solar radiation and dewpoint temperature, (Section 5.2.2.3) and daily observed streamflow provided by Rio Tinto (Section 5.2.2.1) for the period from January 1979 to December 2014. The LSTM models were trained on sequences of [365-k] days of hydrometeorological data prior to a single observed streamflow target, where k is the lead-time for which the model will be used to forecast flows (see Figure 5.2a). Additionally, streamflow data up to the forecast date were used as inputs to the LSTM model, allowing the network to access recent information about the current hydrological state at the time of forecast, similar to how CEQUEAU has access to assimilated initial states. However, this also means that for each forecast lead-time, the LSTM model needs to be retrained with streamflow observations being lagged by the same number of days as the forecast lead-time. For example, for a 3-day lead-time forecast, the LSTM inputs would be a combination of 362 days of observed weather data prior to the forecast day as well as 362 days of streamflow observations lagged by 3 days (see Figure 5.2a for an example of data and periods used to train a 1-day and a 3-day lead-time forecasting model). This ensures that no observed streamflow from the forecast period is used during model training. To train these models, observed data (both weather and hydrometric) are used for all time steps; however, in forecasting, the forecast lead-time days are replaced with actual forecast data and the observed streamflow time series ends on the day of the forecast issue (see Figure 5.2b for an example of application of a 1-day and a 3-day forecasting model). Nine different LSTM models were thus trained, corresponding to each lead-time (one to nine days) of the forecast.

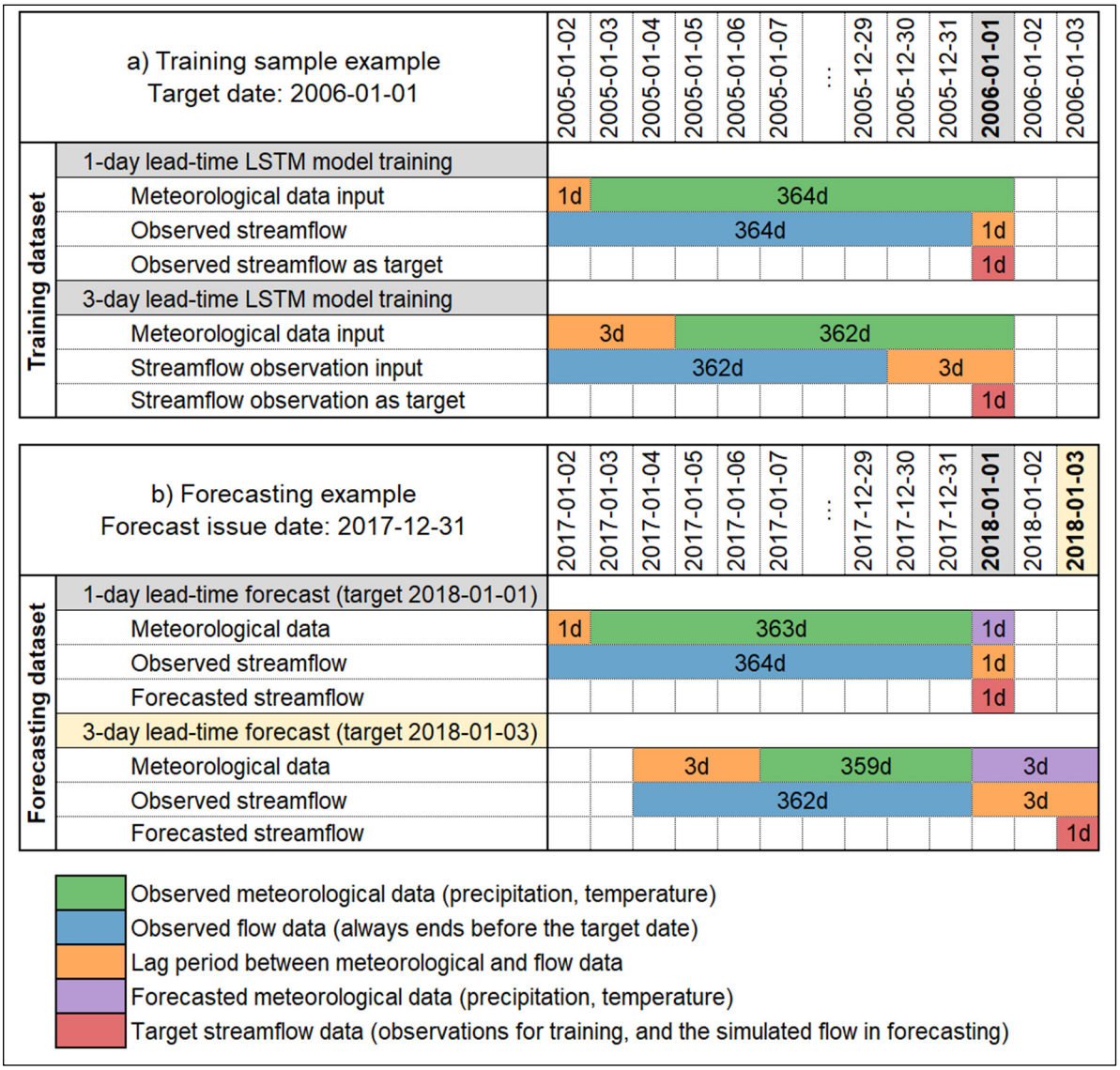


Figure 5.2 Example implementation of a training sample for a 1-day and a 3-day lead-time forecasting model (a) and an example implementation of such forecasting models for 1-day and 3-day lead-time streamflow forecasts (b). Panel (a) presents a single sample, but the process is repeated on the entire dataset available for the training period and the same process is repeated on the validation and testing periods. Panel (b) shows an example for a single issue date, but the process is repeated for other forecast issue dates, combining the outputs of the 1-day to 9-day forecasts for each forecast issue date

The network weights were optimized to maximize model performance over all training sequences. The training set was composed of 70% of the years selected at random from the training period, 15% of the years were kept as the validation dataset to prevent the LSTM from

overfitting during training, and the remaining 15% of the dataset is used as testing data which are used to evaluate the LSTM's robustness on an independent period prior to forecasting.

Data normalization (i.e., scaling of data to ensure data values are within bounds of [0-1]) was applied to all hydrometeorological data using a scaling model calibrated on the training period data only, to ensure no contamination of the training dataset was accidentally introduced. Training and validation were performed using the Tensorflow and Keras libraries in Python, with the Adam optimizer and the KGE metric as the objective function. The model was trained by generating one streamflow value for each [365-k]-day window of training data, and repeated for all days in the dataset, generating a hydrograph one day at a time in multiple batches. The trained model was then used to perform forecasts for the period from January 2015 to December 2019 by combining historical (ERA5 reanalysis dataset) and forecast (ECMWF) data into sequences of [365-k] days, ending on the desired lead-time's meteorological data. One LSTM model was trained for each forecast lead-time due to lagged streamflow observations. The LSTM forecasts were also generated every three days and on the same dates as the CEQUEAU-DA, DA- and OL-forecasts were performed and compared with those of the LSTM.

## 5.4 Results

This section presents the performance assessment of the LSTM model in simulation (Section 5.4.1), the results of the LSTM-based and CEQUEAU-based forecasting for lead-times varying from one to nine days (Section 5.4.2), as well as the results of the DA aspect of forecasting between both approaches (Section 5.4.3).

### 5.4.1 Performance of the LSTM model in simulation

Table 5.1 presents the statistics of the LSTM performance over the training and testing periods. Overall, the LSTM model performs well in simulating daily streamflow as suggested with

KGE values above 0.90 for all lead-times over the training period and above 0.89 for the testing period.

Table 5.1 Description of the LSTM models hyperparameters used for each of the forecast lead-times and training results

| Lead-times (days) | Nb. of LSTM units | Dropout rate | Nb. of training epochs | Batch size | Training KGE | Validation KGE | Testing KGE |
|---|---|---|---|---|---|---|---|
| 1 | 128 | 0.2 | 100 | 128 | 0.97 | 0.99 | 0.98 |
| 2 | 128 | 0.2 | 200 | 128 | 0.98 | 0.99 | 0.98 |
| 3 | 128 | 0.2 | 200 | 128 | 0.95 | 0.95 | 0.95 |
| 4 | 128 | 0.2 | 250 | 128 | 0.95 | 0.91 | 0.94 |
| 5 | 128 | 0.2 | 250 | 128 | 0.92 | 0.94 | 0.91 |
| 6 | 128 | 0.2 | 300 | 128 | 0.93 | 0.91 | 0.92 |
| 7 | 64 | 0.1 | 1000 | 64 | 0.94 | 0.94 | 0.93 |
| 8 | 64 | 0.1 | 1000 | 64 | 0.95 | 0.93 | 0.92 |
| 9 | 64 | 0.1 | 1000 | 64 | 0.90 | 0.92 | 0.88 |

According to the daily hydrograph (Figure 5.3a), although the LSTM model captures the lowest peak flows, a slight underestimation of the highest peak flows is observed, ranging from 5 to 10%. The timing of all peak flows is successfully captured during the training period. As for the testing period, the KGE values are slightly better than those obtained over the training period for day one, and the performance is comparable until day three (Table 5.1). However, for lead-times beyond day four, the LSTM performance over the testing period is slightly lower, even though the KGE values remain very good (KGE $\geq$ 0.88).

In the testing period, the LSTM model slightly underestimates the highest peak flows (3 % on average) while the timing of peak flows is generally well simulated. These results can be seen

in Figure 5.3b for the 1-day lead-time LSTM model. Results for longer lead-times (from 2- to 9-day) show similar skill levels.

Figure 5.4 shows a quantile-quantile plot of the observed and simulated streamflows for each of the nine lead-times during the testing period. A general underestimation of peak flows is seen in most lead-times. Additionally, a small overestimation of the lowest flows can be noted for most lead-times, which can be seen in more detail in the supplementary materials (Figure S5.10).

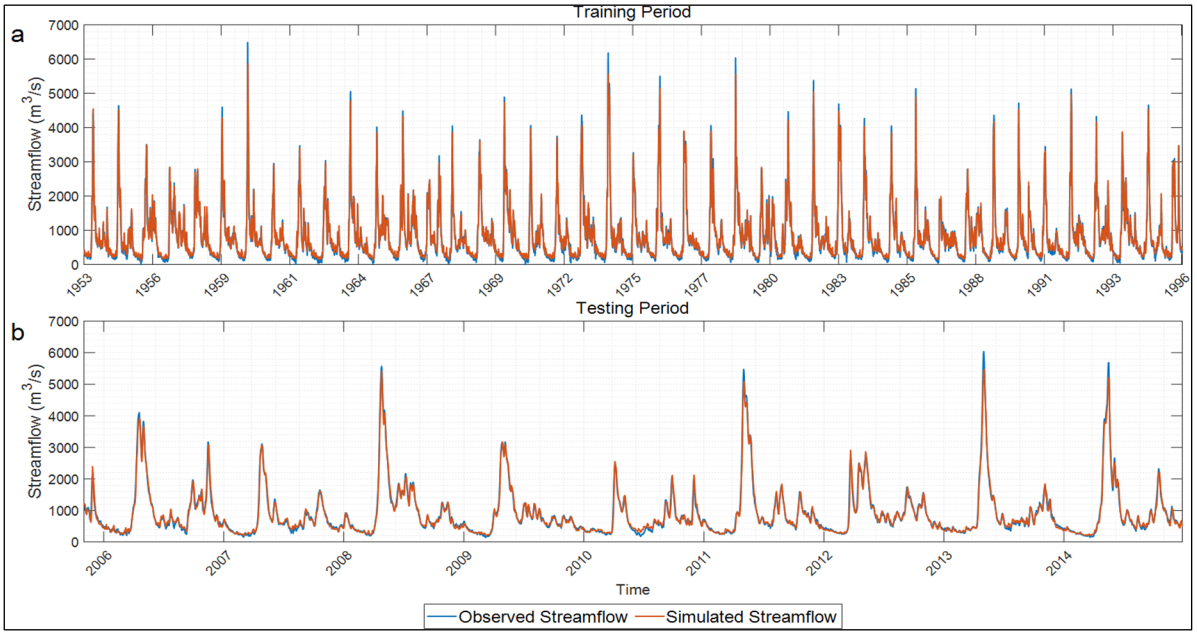


Figure 5.3 Comparison between observed and simulated streamflow from the LSTM models over the 1953-1996 training period (a) and the 2005-2015 testing period (b), using a 1-day lead-time

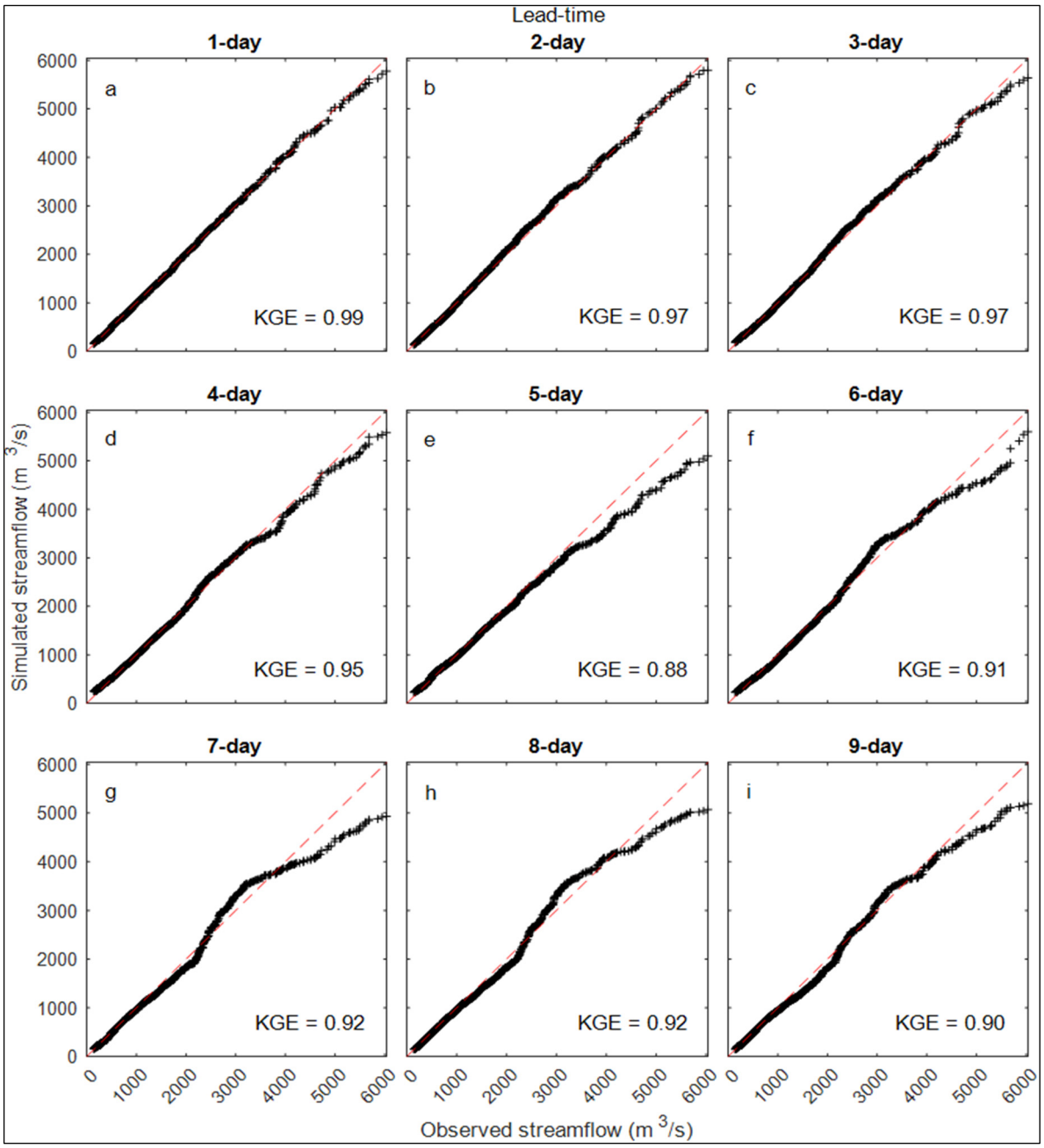


Figure 5.4 Quantile-quantile plot showing observed streamflow against simulated values from the nine LSTM models over the testing period, from 1 to 9 days in lead-time from a) to i). The 1:1 slope (dashed red line) is added for comparison purposes, representing a perfect match between observed and simulated streamflow. Each panel contains 3341 points, i.e., the 10 years of testing data.

### 5.4.2 LSTM and CEQUEAU comparison in forecasting

The main focus of this section is to compare the performance of the LSTM and the CEQUEAU models with data assimilation (CEQUEAU-DA) and without (open-loop, CEQUEAU-OL) when simulating streamflow over the LSJ catchment at the annual and seasonal scales during the testing period.

The skill of ensemble forecasts for one to nine days of lead time is analyzed in Figure 5.5 and Figure 5.6 and Figure 5.7. Figure 5.5 presents the CRPS and MAE scores using data from all years and all seasons as a first step to assess overall model performance. Figure 5.6 and Figure 5.7 present the forecast skill metrics for forecasts issued for each season (i.e., winter: December to March - DJFM; spring: April and May - AM; summer: June, July, and August - JJA; and fall: September, October, and November- SON) for both the CRPS and MAE, respectively. These figures include the results of all forecast issue days within their respective periods (i.e., all forecasts generated during the summer days - 122 initial dates of forecast - are represented in the summer boxplots in Figure 5.6 and Figure 5.7). Lower CRPS and MAE values indicate more accurate forecasts.

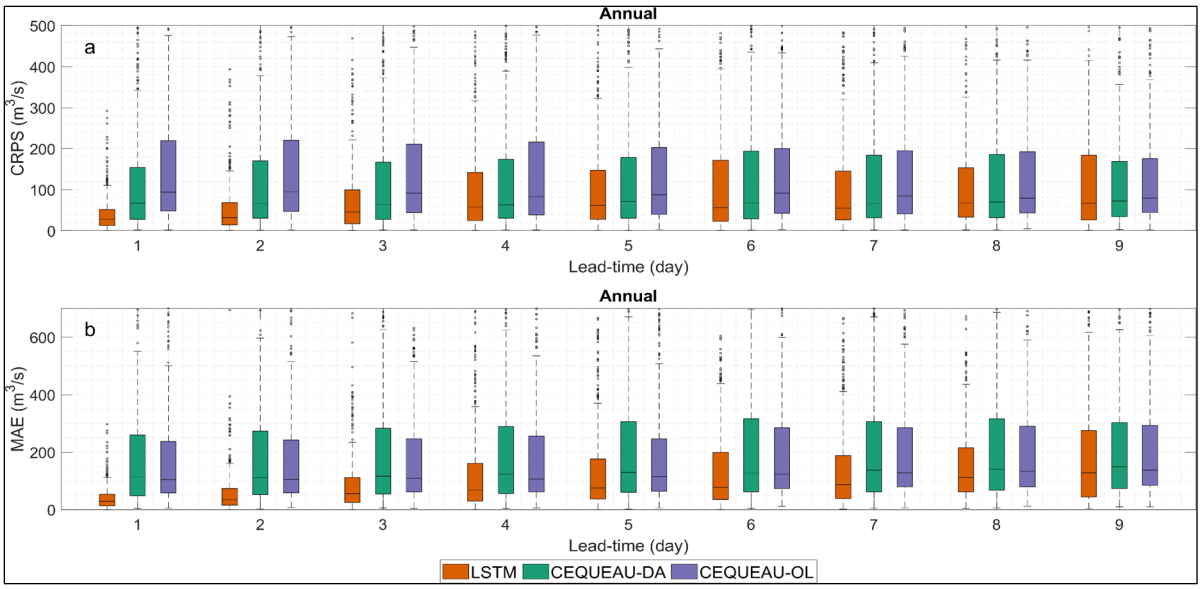


Figure 5.5 CRPS (a) and MAE (b) of the annual streamflow ensemble forecasts for the LSTM model (orange), CEQUEAU-DA (DA; green), and CEQUEAU-OL (OL; purple) over the 2015-2019 forecasting period. Each boxplot contains 536 forecasts, corresponding to one forecast every three days over the study period. The center horizontal line in each boxplot represents the median, box edges represent the 25th and 75th percentiles, and whiskers represent the extreme values not considered as outliers. Dots outside of the whiskers are outliers

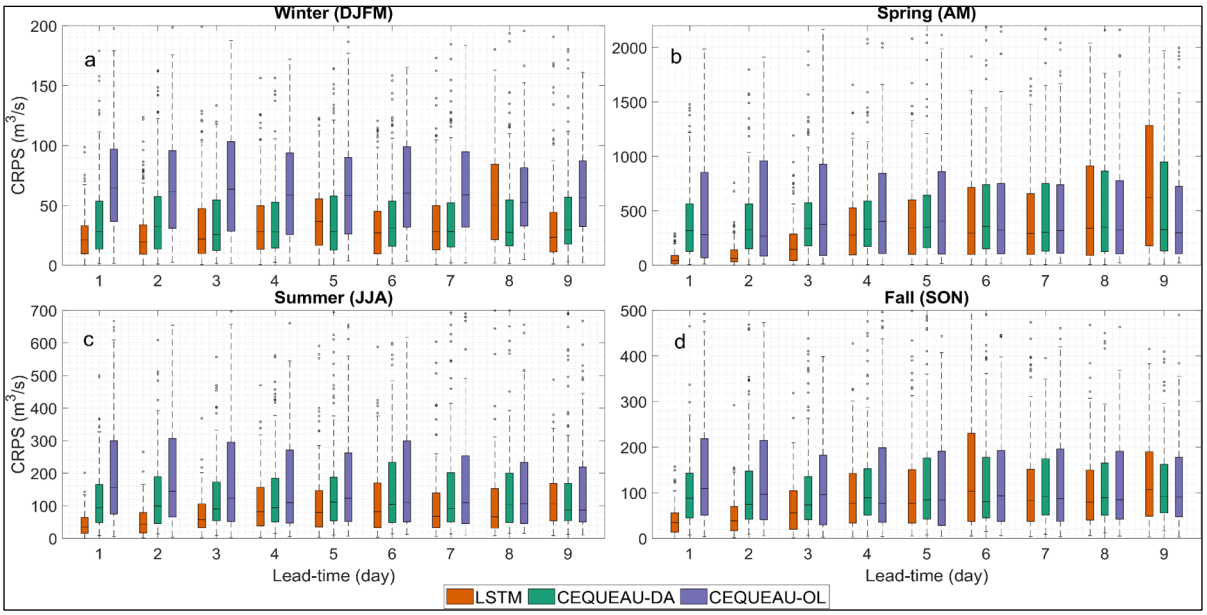


Figure 5.6 CRPS of the seasonal streamflow ensemble forecasts for the LSTM model (orange), CEQUEAU-DA (green), and CEQUEAU-OL (orange) over each season of the 2015-2019 forecasting period: winter: (December to March; DJFM - a), spring (April and May; AM - b) summer (June to August; JJA - c), and fall (September to November; SON - d). The number of points in each boxplot represents the number of issued forecasts for that season, equal to 191, 101, 122, and 122 for Winter, Spring, Summer, and Fall, respectively. Note that the y-axis ranges are different in all panels due to the large differences between seasons and some outliers are not shown for clarity's sake

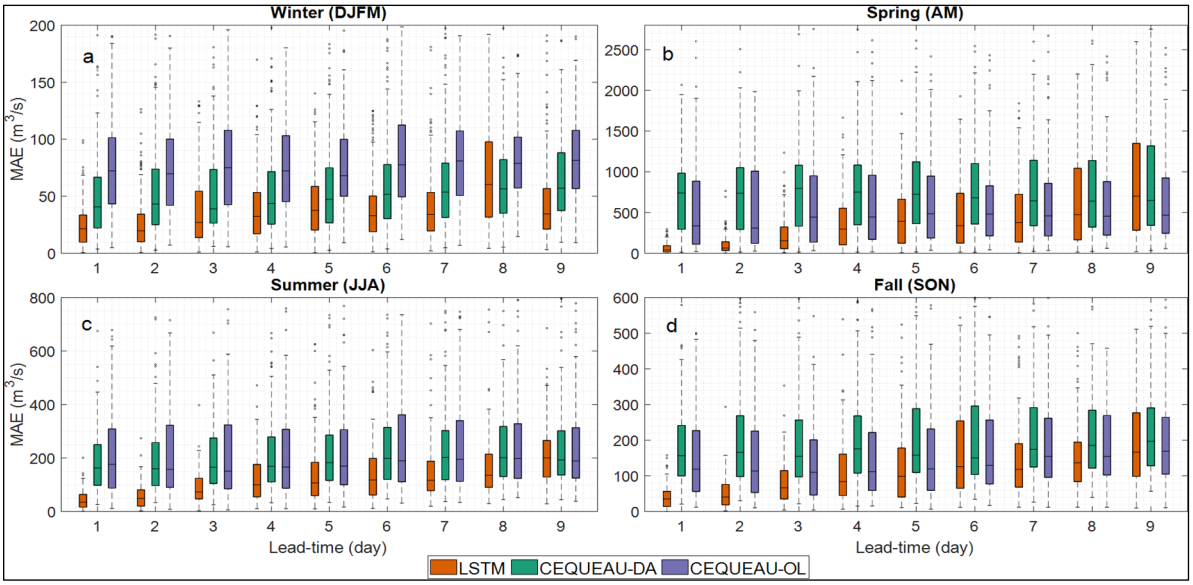


Figure 5.7 MAE of the seasonal streamflow ensemble forecasts for the LSTM model (orange), CEQUEAU-DA (green), and CEQUEAU-OL (purple) over each season of the 2015-2019 forecasting period: winter: (December to March; DJFM - a), spring (April and May; AM - b) summer (June to August; JJA - c), and fall (September to November; SON - d). The number of points in each boxplot represents the number of issued forecasts for that season, equal to 191, 101, 122, and 122 for Winter, Spring, Summer, and Fall, respectively. Note that the y-axis ranges are different in all panels due to the large differences between seasons and some outliers are not shown, for clarity's sake

The annual results indicate that LSTM performs better than both CEQUEAU-DA and CEQUEAU-OL for both CRPS and MAE values up to the 8-day lead-time (Figure 5.5). On average, the median LSTM performance is 22% (42%) better for CRPS (MAE) compared to CEQUEAU-DA, and 37% (37%) better for CRPS (MAE) compared to CEQUEAU-OL across all lead-times.

The LSTM model outperforms CEQUEAU during the first three days of the forecast, as seen in the third quartile of the LSTM CRPS which is inferior (better) to the median CRPS of the CEQUEAU forecasts for the first two days, and a similar trend can be observed for MAE. This is supported by a non-parametric Wilcoxon test for median equality, which shows the results on an annual basis for both skill scores, as depicted in Table 5.2a.

For days 4-9, the LSTM model still performs better than CEQUEAU, but by a smaller margin. The MAE results, on an annual scale, show that LSTM has significantly lower MAE values than CEQUEAU (using both DA and OL methods) for the nine lead-times (Table 5.2b). The CRPS results also support these findings, as shown in Figure 5.4. However, after the 3-day lead-time, there is no significant difference between LSTM and CEQUEAU-DA for the CRPS.

Additionally, CEQUEAU-DA generally provides better CRPS and MAE values than the OL scenario when looking at forecast quality over the entire year, which is expected. It should also be noted that the LSTM displays progressively wider spread (i.e. wider interquartile range) in CRPS and MAE as lead-time increases (for example, interquartile ranges of CRPS for days 1, 3, 5 are 39, 82, and 119 $m^3/s$, respectively). This is likely attributed to the fact that the LSTM model has the exact same starting conditions for every member of the ensemble during a forecast, and only the forecast weather can contribute to the variability, as will be discussed further.

Table 5.2 Results of a Wilcoxon rank sum statistical test with a significance level of 1% for CRPS (a) and MAE (b) results between the forecasts generated by the three models used in this study (LSTM, CEQUEAU-DA and CEQUEAU-OL). A value of H=0 indicates equal medians between the forecasts of the groups defined in each column, whereas a value of H=1 indicates that the null hypothesis is rejected, indicating different medians for the two groups. Performance is evaluated per lead-time

(a) CRPS

| | H - Result of the hypothesis test | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Winter | | Spring | | Summer | | Fall | | Annual | |
| Lead-time | LSTM vs. DA | LSTM vs. OL | LSTM vs. DA | LSTM vs. OL | LSTM vs. DA | LSTM vs. OL | LSTM vs. DA | LSTM vs. OL | LSTM vs. DA | LSTM vs. OL |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 3 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 4 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 5 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| 6 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 |
| 7 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| 8 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| 9 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |

(b) MAE

| | H - Result of the hypothesis test | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Winter | | Spring | | Summer | | Fall | | Annual | |
| Lead-time | LSTM vs. DA | LSTM vs. OL | LSTM vs. DA | LSTM vs. OL | LSTM vs. DA | LSTM vs. OL | LSTM vs. DA | LSTM vs. OL | LSTM vs. DA | LSTM vs. OL |
| 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 3 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| 4 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 |
| 5 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 |
| 6 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 |
| 7 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 |
| 8 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 |
| 9 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

### 5.4.3 Forecast performance evaluation

To evaluate the performance of the LSTM model in comparison to traditional hydrological models, ensemble streamflow forecasts generated with the LSTM model are compared to those produced by CEQUEAU-DA and CEQUEAU-OL models for a 9-day lead-time, the maximum available in the weather forecast data used in this study. The performance is evaluated for each season by selecting a typical forecast event and using the forecast issue date with the median streamflow observation of that season to ensure representativeness (Figure 5.8). Ensemble forecasts using the three methods (LSTM, CEQUEAU-DA and CEQUEAU-OL) are then evaluated for these selected events.

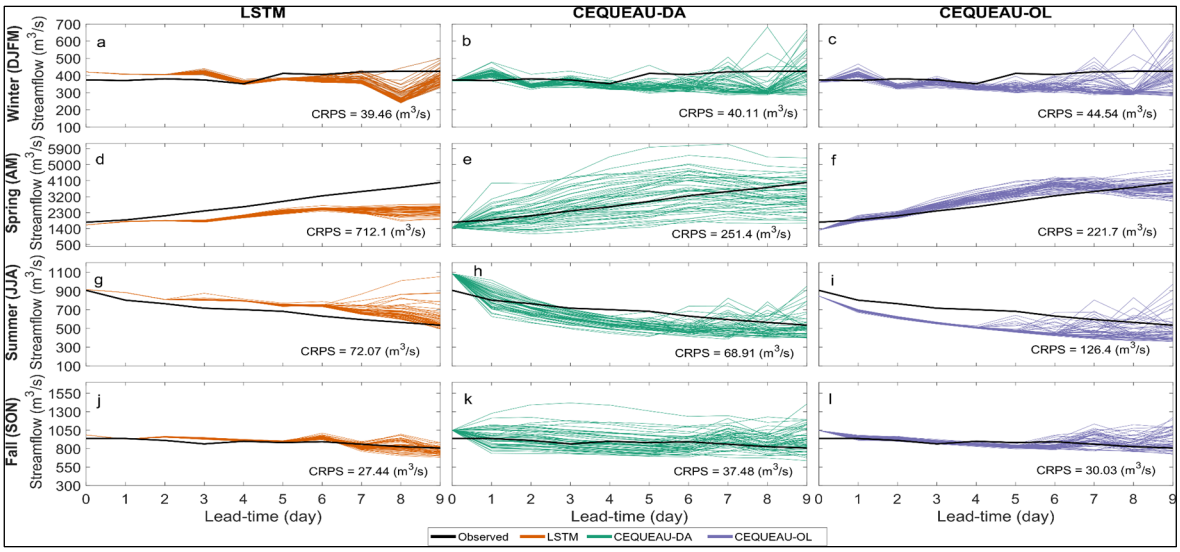


Figure 5.8 Forecasted streamflow ensembles for each season (winter: a, b, c; spring: d, e, f; summer: g, h, i; fall: j, k, l) as generated by the LSTM model (left column; a, d, g, j), CEQUEAU-DA (middle column; b, e, h, k), and CEQUEAU-OL (right column; c, f, i, l) over a 9-day lead-time. For each row (i.e., season), the forecast date chosen for display is that which corresponds to the day where the observed flow is the median value of all observations for that season. The exact dates are Jan-26-2016 (Winter), Apr-26-2017 (Spring), Jul-30-2016 (Summer) and Nov-21- 2015 (Fall)

The results in Figure 5.8 indicate the added value of the LSTM model to the overall quality of ensemble forecasts in different seasons for various lead-times. In winter, LSTM provides an

almost unbiased forecast up to approximately the 6-day lead-time, compared to the CEQUEAU forecasts. The CRPS for the LSTM forecast is also lower than that of the CEQUEAU forecasts over the entire period due to the ensemble having a lower spread. The initial states are well simulated by all three forecasting models and are not a factor in the bias related to initial conditions.

In spring, forecasts vary significantly depending on the forecast model. The LSTM has much less variability than the CEQUEAU-OL forecast, which is in turn less variable than the CEQUEAU-DA implementation. The latter has more variability primarily due to the assimilation of initial states, making it more sensitive to input meteorological data variations. Additionally, the LSTM forecasts for the first days display less bias than the hydrological model counterparts, owing to the integration of recent streamflow as inputs. It can also be seen that up to a lead-time of approximately five days, the LSTM forecast shows better skill (smaller ensemble error) than the other models. However, in the following days, the small spread and increasing bias of the LSTM ensemble members heavily penalized the CRPS score, making it worse overall than the CEQUEAU implementations. This is also seen in Figure 5.6b, where the LSTM generally produces better forecasts for shorter lead-times only. In summer, LSTM decreases the spread of the ensemble forecasts over the 9-day lead-time, allowing it to get closer to the observed streamflow compared to both the CEQUEAU-DA and CEQUEAU-OL ensembles except during the last forecast days, contributing to the improvements in CRPS (see Figure 5.6c). In fall, the comparison of LSTM-ensemble forecasts with the CEQUEAU-DA and CEQUEAU-OL ensembles shows that LSTM generates reliable forecasts that encompass the observation with very little spread for the entire forecast duration. The CEQUEAU-DA, however, has initial conditions that are more representative of the current state and are more saturated (i.e., more reactive) than CEQUEAU-OL for that given period.

Overall, the main difference between the LSTM and CEQUEAU-based forecasts is that the LSTM is more confident in its forecasting, generating forecasts with less spread. When the forecast is unbiased, the forecast skill is better than that of the hydrological models. However,

when the LSTM model generates a biased forecast, the low spread makes all 50 members biased, thus increasing the CRPS and MAE error metric values.

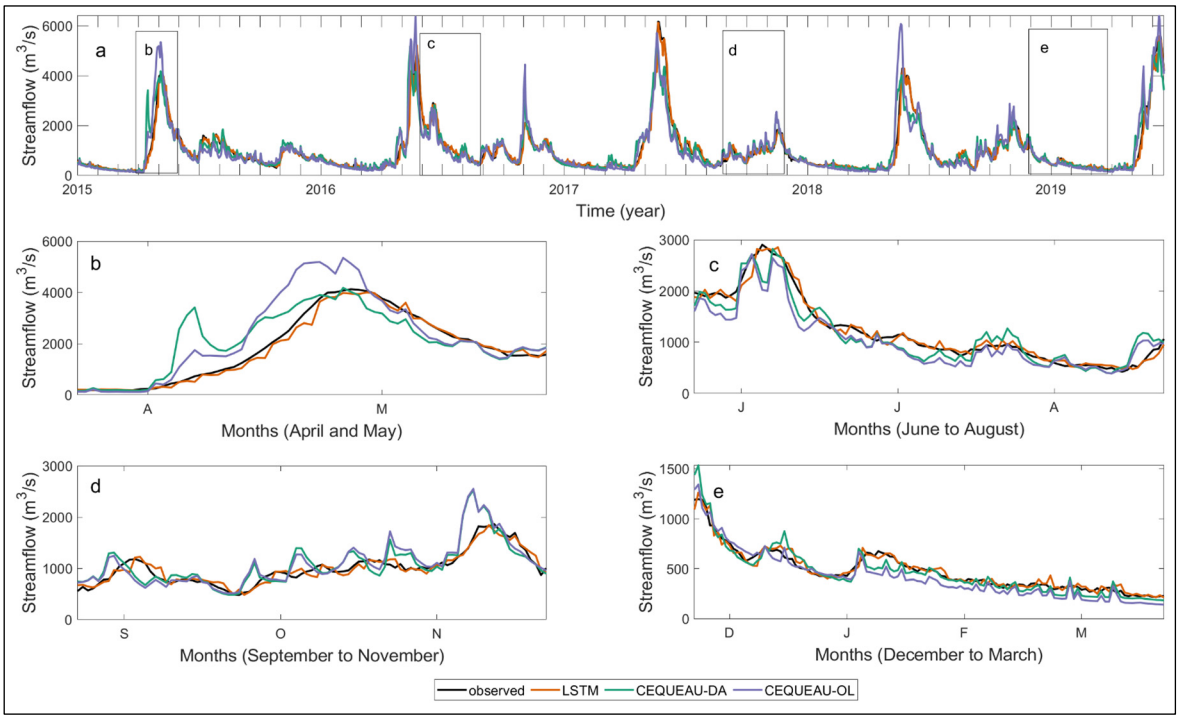


Figure 5.9 Hydrographs generated from successive 1- to 3-day lead-time streamflow forecasts averaged over the 50 members for the LSTM (orange), CEQUEAU-DA (green), and CEQUEAU-OL (purple) models and observations (black) over the period between January 2015 and May 2019 (a). The focus is placed on the individual seasons of spring (b), summer (c), fall (d) and winter (e)

The forecasted hydrographs of each model are compared in Figure 5.9. The results illustrate that the LSTM is more accurate at shorter lead-times (1-day to 3-day) than the CEQUEAU-DA and CEQUEAU-OL forecasts throughout the year. Similar results are found for the 4- to 6-day and 7- to 9-day forecasts presented in the supplementary materials Figure S5.11 and Figure S5.12. The longer lead-times demonstrate that the LSTM errors are primarily caused by temporal shifts, while the CEQUEAU errors are mainly attributed to amplitude errors, but with better timing.

## 5.5 Discussion

The following section presents a comparison of the development and performance of the CEQUEAU (DA and OL) and LSTM models for forecasting streamflow (Section 5.5.1 ). The importance of DA and its effect on model performance is then discussed (Section 5.5.2). Finally, an evaluation of the impact of using observed streamflow as a predictor on the LSTM model performance is made (Section 5.5.3).

### 5.5.1 Comparison between CEQUEAU and LSTM

The CEQUEAU and LSTM models have some similarities and key differences when it comes to forecasting streamflow. CEQUEAU-DA initiates forecasts with an improved estimate of initial states, resulting in a more sensitive model from the first lead-time. On the other hand, the LSTM model is initialized deterministically using past weather and streamflow observations, resulting in a lack of uncertainty in the 1-day lead-time (Figure 5.8). Both models are trained on historical data, but CEQUEAU is calibrated over a historical time period and used to generate forecasts based on incoming weather forecasts, while the LSTM model is trained on a set of two inputs (precipitation and temperature) and one output (streamflow). The length of the training is always 365 minus the lead-time duration (i.e., 364 days descending to 356 days). For forecasting, the same size and data standardization are used, so the very end data of the inputs (i.e., the same as the desired lead-time) is used as input for the weather forecast, while the streamflow forecast is generated for the desired lead-time.

CEQUEAU-DA produces a wider spread of forecast results compared to the LSTM model. Both LSTM and CEQUEAU (both DA and OL) display CRPS and MAE errors worsening with increasing lead-times. This is likely due to the decreasing reliability of weather forecasts as lead-times increase, resulting in less skillful streamflow forecasts from both LSTM and CEQUEAU. Overall, the results indicate that LSTM provided better forecast results compared

to the CEQUEAU model, while also requiring simpler implementation and no DA. However, the LSTM model has the limitation of needing to be trained with multiple times for different lead-times. Furthermore, while the LSTM model may perform well on a single catchment as seen with the LSJ, it might also benefit from training over several basins to be robust and efficient at capturing extreme events, such as in Kratzert et al. (2018). In this study, LSTM is only evaluated for one catchment, which limits the generalizability of the findings. Finally, it is important to note that the LSTM trained in this study used a wide array of combinations of ERA5 meteorological time series. However, the most robust model tested only made use of precipitation and temperature. Adding more variables made the training more difficult and either converge to worse values or not converge at all. It could be possible to improve on these results by adding more types of data (such as snowpack depth, more complex meteorological variables and other lagged variables such as climate indices data) and finding LSTM model structures and hyperparameters that can make use of the extra input data, however this was not successfully implemented in this study.

### 5.5.2 On the necessity of performing DA

Hydrological models have evolving states that can drift from the real hydrologic state if left unchecked during long simulations. To correct this drift and ensure the hydrological model is as close as possible to the real hydrological state, DA can be performed. The impact of DA on forecasting performance can be evaluated by comparing the results of a model setup without performing DA (i.e, OL) to a setup that constantly assimilates newly available data. DA typically improves forecast skill (i.e., in terms of accuracy and reliability), particularly on shorter lead-times.

Although LSTMs do not typically implement DA directly, they do have a recurrent state that allows for the ingestion of near real-time observations (Nearing et al., 2022). In this study, DA was not performed directly. Instead, the integration of observed streamflow at the forecast issue date was used as a proxy. This means that the result of the previous simulation has no impact

on the current forecast and the LSTM breaks continuity in the forecasting stage. Hydrological models, on the other hand, ensure the forecast is done in one single run, which is an advantage. The LSTM used in this study therefore trades continuity for not having to resort to the complex step of DA given the number of memory states in the LSTM. By providing observed streamflow in the training and forecasting steps, the LSTM can learn to minimize initial state errors, similar to the DA step in a hydrological model using observed streamflow to adjust its initial states.

Both the hydrological model with DA and the LSTM model provided skilled streamflow simulations and forecasts. This study found that for the catchment studied, the LSTM outperforms the hydrological model for short-term forecasts (up to 5-9 days lead-time depending on the season) when using the CRPS and MAE metrics. Figure 5.4a and Figure 5.4b showed that the LSTM performed better than the CEQUEAU-DA and CEQUEAU-OL models for all lead-times with a significant improvement in forecast skill. One advantage of the LSTM is that it does not require a DA scheme, which can lead to less computational effort and easier implementation.

### 5.5.3 Impacts of using observed streamflow as a predictor

In this study, the LSTM used observed streamflow up to the forecast issue date to provide information on the current hydrological state. The use of observed streamflow as a predictor is not a new concept in hydrology; in fact, it has been a useful proxy for short-term forecasts over the last decade (Cloke & Pappenberger, 2009). Different studies have applied various methods such as using historical streamflow observations directly to estimate future outcomes (e.g., Rajagopalan et al. (2010)), applying regression models (Bogner et al., 2016; Hopson & Webster, 2010; Seo et al., 2006), using autoregressive models (e.g., ARMA, ARMAX, GARCH (Amiri, 2015; Zhang et al., 2015) and using artificial neural networks (Coulibaly et al., 2000; Machado et al., 2011).

Recent studies have focused on using LSTM for its ability to simulate hydrological processes and provide high-quality streamflow simulations and forecasts from historical weather data. In this study, streamflow was added as an input to the LSTM. By training the model to combine the information from observed streamflow and weather up to the forecast issue date, as well as forecasted weather, the LSTM is able to make accurate streamflow forecasts for shorter periods. However, as the lead-time increases, the impact of observed streamflow diminishes because the streamflow observations are not available past the forecast issue date. This is reflected in the progressive widening of the CRPS and MAE values as lead-times increase (see Figure 5.4 to Figure 5.6). To improve performance for longer lead-times, the LSTM must be trained independently for each lead-time, with the observed streamflow lagged by the appropriate number of days. This makes the training process more labor-intensive and the model is less able to rely on the observed streamflow for predicting the forecasted streamflow, which results in progressively worse CRPS and MAE scores, similar to the forecasts issued using the hydrological model.

Additional tests were conducted to overcome the need for training one LSTM per lead-time in order to improve forecasting performance. One test involved training the LSTM using only historical weather data as the input, rather than incorporating observed streamflow data. This approach was found to provide good results for only two days of lead-time. However, for longer lead-times, the LSTM performed worse than the hydrological models (CEQUEAU-DA and CEQUEAU-OL). This could be due to the fact that the forecast data from the ECMWF could have different statistical properties than the observations used to train the LSTM, leading to a bias in the forecast. Incorporating observed streamflow data may help to minimize this effect by reducing the weight of the weather component in the forecasting chain.

Another test involved using an LSTM trained on a 1-day ahead forecast to simulate flows for the next day, and then applying the model sequentially to obtain flows that are always lagged by only 1 day for each forecast lead-times. This approach also resulted in poor performance for lead-times longer than two days. This could be due to errors propagating with each iteration,

as small errors in the streamflow for day 1 are then used as input for the second day and so on. Training an LSTM for each timestep individually helps to prevent these errors by addressing any systematic biases at longer lead-times during the training process and from within the weather forecast itself.

## 5.6 Conclusion

This study evaluates the potential of the LSTM in simulating and forecasting streamflow of the Lac-Saint-Jean catchment in Canada. It compares the LSTM to the operational CEQUEAU model used by Rio Tinto for this catchment, which is set up in open-loop mode (CEQUEAU-OL) and with a DA scheme (CEQUEAU-DA), which is used as a benchmark. The main findings of this study are as follows:

(1) The LSTM achieves good performance in the training and testing periods for lead-times up to 9 days with a KGE higher than 0.88.
(2) The LSTM provides more skillful ensemble forecasts compared to CEQUEAU-OL and CEQUEAU-DA, as CRPS and MAE results show lower values for the LSTM, all percentiles considered.
(3) The LSTM forecasts display tighter spreads than the CEQUEAU-based forecasts, likely due to the strong influence of the observed streamflow from the previous days used as a predictor, as opposed to the DA implementation that contains uncertainty.
(4) The LSTM eliminates the need for integrating a DA process, typically required by traditional hydrological models, while still providing high-quality forecasts.

The findings of this study have highlighted the advantages, limitations, and specific evaluation of the LSTM performance in streamflow forecasting for all seasons. Overall, this study shows that LSTM is a promising model for forecasting short-term streamflow, and confirms previous findings in other regions and catchments. It is likely that more advanced neural networks and data integration strategies will lead to even more significant improvements. However, this study demonstrates that, in this snow-dominated North American catchment, LSTM models

can provide short-term streamflow forecasts with better accuracy than those generated by more complex distributed hydrological models.

## 5.7 Acknowledgements

This study was partially funded by the Natural Sciences and Engineering Research Council of Canada (NSERC) under the Collaborative Research and Development grant CRDPJ-522126-17. The authors would like to thank Rio Tinto for sharing their hydrometeorological data on the Lac-Saint-Jean catchment. The authors are also grateful to the European Center for Medium-Range Weather Forecasts (ECMWF) for providing access to the historical forecast data from their MARS computing and archiving facilities. In this study, the ERA5 reanalysis dataset produced by Hersbach et al., (2018) was used. It has been downloaded from the Copernicus Climate Change Service (C3S) Climate Data Store: https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-single-levels?tab=overview. The base map in Figure 5.1 was created using ArcGIS® software by Esri. ArcGIS® and ArcMap™ are the intellectual property of Esri and are used herein under license. Copyright © Esri. All rights reserved. For more information about Esri® software, please visit www.esri.com.

## 5.8 Supplementary material

### 5.8.1 Supplementary Materials – S1

#### 5.8.1.1 DA procedure for CEQUEAU

Sabzipour et al. (2022) (*underreview article* – CHAPTER 4 of this document) performed a DA study to evaluate the best hyperparameters for implementing DA on the LSJ catchment. They

used EnKF as the assimilation method, identical to the one used in this study. DA was done seasonally, so hyperparameters (two inputs: mean daily temperature uncertainty, total daily precipitation uncertainty, and one output: observed daily streamflow uncertainty) and 'selected state variables' (vadose zone, saturated zone, snowpack) were recommended for each season of the year. Seasons were divided according to the months of December to March for winter, April and May for spring, June to August for summer, and September to November for fall. Table S5.3 presents recommended recipes for hyperparameter and state variables. It shows the hyperparameters in terms of hydrometeorological variable uncertainty should be chosen for inputs and outputs, and which state variable(s) should be updated to extend the forecast skill further.

Table S5.3 Summary of the recommended DA recipe for each season (adapted from Sabzipour et al., 2022). The observations are perturbed by adding Gaussian (i.e., for temperature and streamflow) or Gamma (i.e., for precipitation) noise from the error distribution (sampling errors). Uncertainty value (Hyperparameter) was used as the mean of an error distribution (Gaussian for temperature and streamflow, Gamma for precipitation) for sampling errors

| | | Winter | Spring | Summer | Fall |
|---|---|---|---|---|---|
| Hyperparameter | Temperature uncertainty (standard deviation of the gaussian error distribution during sampling) | 3°C | 5°C | 7°C | 8°C |
| | Precipitation uncertainty (percentage of the observed precipitation value used to scale the $\alpha$ and $\beta$ parameters of the gamma error distribution) | 10% | 10% | 10% | 10% |
| | Streamflow uncertainty (percentage of the observed streamflow value used as the standard deviation of the gaussian error distribution) | 5% | 5% | 5% | 5% |

| | | | | |
|---|---|---|---|---|
| State variables to be updated | vadose zone | vadose zone | vadose zone | vadose zone, snowpack |

### 5.8.2 Supplementary Materials – S2

#### 5.8.2.1 Quantile-quantile plots of low-flow simulations

Figure 5.4 showed the entire distribution of simulated and observed streamflow data for the entire period. To better show the lower end of the distribution, where there is a high density of data points, Figure S5.10 shows more details on the lower end of the distribution, i.e., only for observed flows below 1000 $m^3/s$. Overestimation of low flows is better displayed here, namely for lead-times 3 to 6 days.

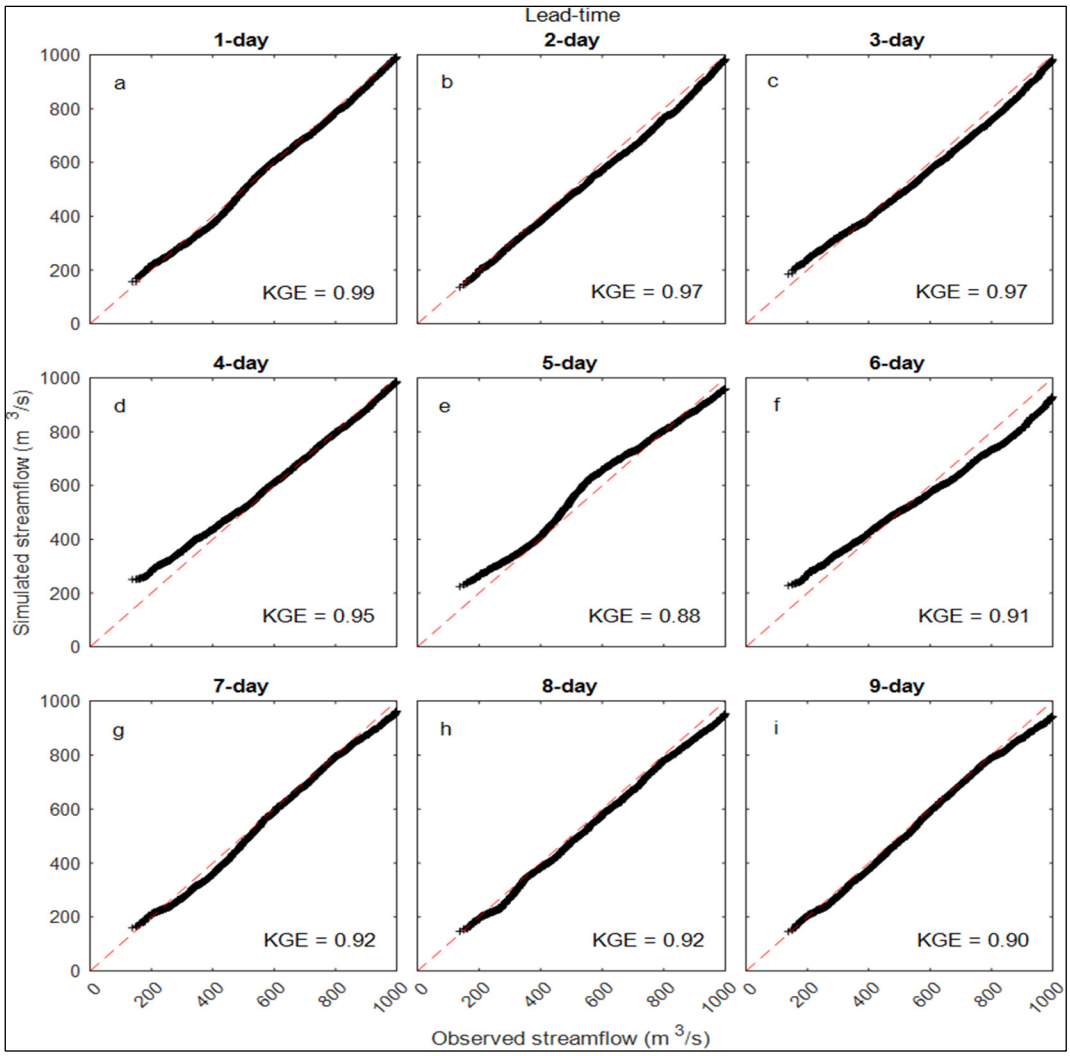


Figure S5.10 Quantile-quantile plot showing the testing period performance of the 9 LSTM models (black markers), labeled 1-9 according to the lead-time used in training. The 1-1 line (dashed red line) is added to ease comparison. This figure shows the lower portion of the distribution shown in Figure 5.4

### 5.8.3 Supplementary Materials – S3

Figure S5.11 and Figure S5.12 represent simulated hydrographs for longer lead-times than those in Figure 5.9. They are constructed from the 4- to 6-day and 7- to 9-day lead-time streamflow forecasts, respectively. The hydrographs consist of successive 3-day forecasts,

generated every 3 days, at the same time as the CEQUEAU forecast is issued after its assimilation. The inserts show subsets of the hydrographs for spring (b), summer (c), fall (d), and winter (e). These hydrographs represent the mean of the 50-member ensemble forecasts generated by CEQUEAU in open-loop (OL) mode, CEQUEAU in DA (DA) mode, and the LSTM.

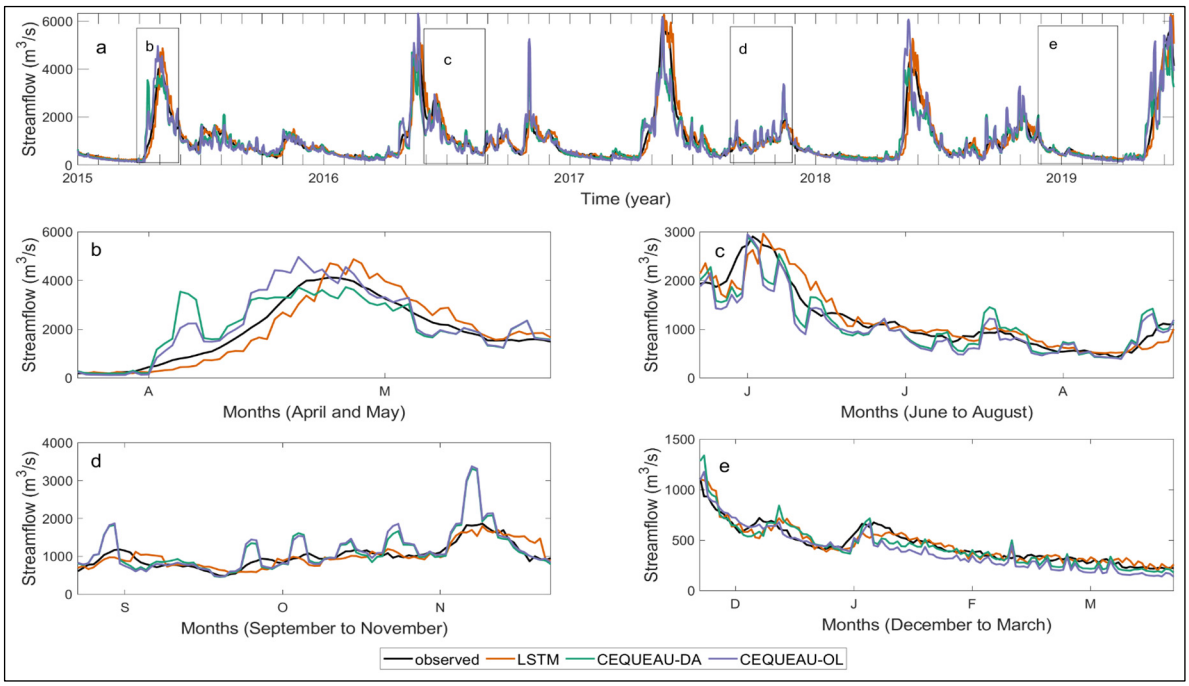


Figure S5.11 Hydrographs constructed from the 4- to 6-day lead-time streamflow forecasts generated every 3 days which are the days where CEQUEAU-DA assimilates data. The hydrographs are thus composed of successive 3-day forecasts, generated every 3 days. Inserts represent subsets of the hydrograph (a) for spring (b), summer (c), fall (d) and winter (e). All hydrographs represent the mean hydrographs from the 50-member ensembles generated with CEQUEAU-OL, CEQUEAU-DA and the LSTM. This figure is the same as Figure 5.9 but using lead-times of 4 to 6 days concatenated to generate a daily hydrograph

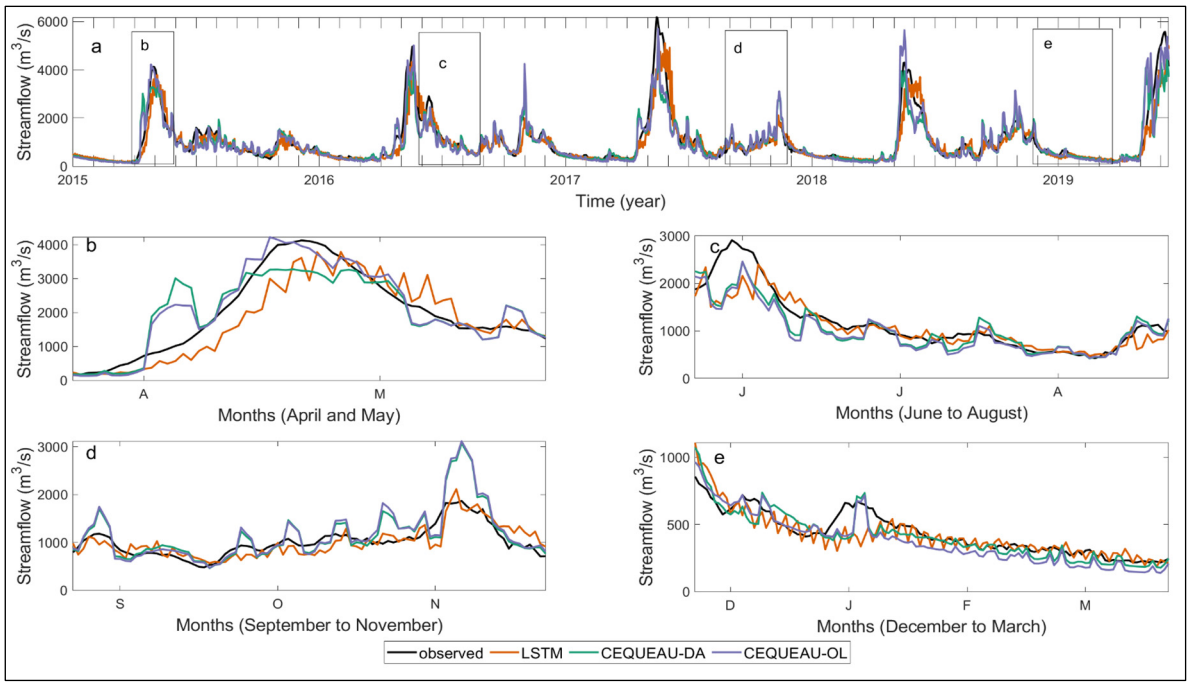


Figure S5.12 Hydrographs constructed from the 7- to 9-day lead-time streamflow forecasts generated every 3 days which are the days where CEQUEAU-DA assimilates data. The hydrographs are thus composed of successive 3-day forecasts, generated every 3 days. Inserts represent the mean hydrographs from the 50-member ensembles generated with CEQUEAU-OL, CEQUEAU-DA and the LSTM. This figure is the same as Figure 5.9 but using lead-times of 7 to 9 days concatenated to generate a daily hydrograph

# CHAPTER 6

# GENERAL DISCUSSION

The main objective of this research was improving streamflow prediction quality, namely in terms of accuracy and reliability, over multiple lead-times. Methods were proposed to tackle this objective in separate sub-objectives. As a reminder, the secondary objectives are as follow:

1) Improving long-term forecasts (i.e., multiple months lead-time) using ESP member filtering.
2) Improving short-term forecasts (i.e., up to 10 days) using distributed hydrological model DA.
3) Evaluating the ability of data-driven models (LSTM) on short-term forecasting while using proxies of DA.

This section summarizes the findings of the studies on these three sub-objectives (sections 6.1, 6.2, 6.3 ) and presents an overall analysis on how these objectives were attained and what has been learned through this research (section 6.46.4). Following this, some limitations of this work and some recommendations for further research are highlighted (section 6.5)

## 6.1 Improving long-term forecasts using ESP member filtering

The accuracy and reliability of streamflow forecasting for periods ranging from weeks to months largely depend on our understanding of the phenomena that govern weather patterns in a region over longer periods than just a few days or weeks. Generating a long-term streamflow forecast, which spans a period of approximately more than two weeks, is a challenging task. The key to success in this task is identifying predictors, such as climate indices, that can provide insight into weather patterns during the forecast period (Devineni et al., 2008). The problem of long-term streamflow forecasting has been approached in several ways in the literature, including filtering and dressing ensemble scenarios (Li et al., 2017; Troin et al., 2021), post-processing (Seo et al., 2006; Verkade et al., 2013), and the use of predictors

such as large-scale climate indices or conditioning on the period prior to the forecast date (Bradley et al., 2015; Hao et al., 2018; Sohrabi et al., 2021).

The first part of this study examined filtering techniques as a way to improve long-term streamflow forecasts. Filtered ensembles are streamflow forecasts that are more likely to occur based on the similarity of hydroclimate conditions before the forecasting date and during the same period in historical data. Filtering techniques can remove unwanted ensemble members by selecting members based on climate indices (Lamb et al., 2011; Najafi et al., 2012; Soukup et al., 2009). Other studies have investigated the use of predictor periods and large-scale climate indices as important factors in predicting streamflow (Shams et al., 2018; Sohrabi et al., 2021; Werner et al., 2004; Zhao & Brissette, 2022). There are several statistical methods that can be used to predict streamflow, such as Bayesian methods (Bradley et al., 2015; Robertson & Wang, 2012), regression methods (Kroll et al., 2004), and analog methods (Hemri & Klein, 2017; Yang et al., 2020). CTs and simple analog methods can be used to test why some ensemble members are selected while others are not.

Reducing the ensemble size can improve forecast skill, but identifying which members to remove optimally is a challenge (Ferro et al., 2008). In this study, we aimed to measure the forecast skill in hydrometeorological historical data for long-term streamflow forecasting using a proposed GA-based method. We tested the CT method (Chandimala & Zubair, 2007; Habets et al., 2004; Jeong & Kim, 2005; Maity & Nagesh Kumar, 2008) and the analogy method (Bellier et al., 2016; Koutsoyiannis et al., 2008; Svensson, 2016; Yao & Georgakakos, 2001) to determine if they could improve streamflow prediction by reliably removing less useful scenarios. We compared more probable scenarios with less probable ones to identify possible indices and find significant differences. However, the statistical characteristic used for these methods was the average streamflow in the predictor period, which was not representative enough to confidently determine which ensemble members lead to lower forecast skill when compared to using all ensemble members (see Figure 3.6 and Figure 3.7). While the CT and analogue methods showed some skill, they did not consistently improve the forecasts.

Unfortunately, we did not achieve our objective of identifying which ensemble members to remove optimally. While the GA method proved to be valuable in determining the highest potential forecast skill that can be achieved using an ESP ensemble, the subsequent analysis did not reveal any reliable patterns to identify or predict which members should be used in actual forecasting situations.

Despite limitations, the proposed GA filtering method has several potential applications for the hydrological forecasting community. Firstly, it can quantify the unrealized skill in the historical data, helping identify the potential of the data in forecasting future streamflow. Secondly, it can sort out basins according to their unrealized potential, and highlight more complex basins with greater potential, as seen in Figure 3.3 and Figure 3.4. For example, the BC catchment loses more members than the QC catchment for the same decrease in CRPS (200 $m^3$/sec), indicating that BC's climate is more predictable than QC's. This is compounded by the fact that the BC catchment has higher overall flows and should thus need a smaller reduction of the ensemble members to achieve this same drop in CRPS. Lastly, it is possible to improve forecast skill beyond the potential of historical data using methods such as neural networks (like LSTM), and the GA method could serve as a benchmark for evaluating these methods. Therefore, the proposed GA method can provide valuable insights and help in improving long-term streamflow forecasting.

However, it is recommended to test the GA filtering method with other statistical characteristics and more complex (multivariate, non-linear, etc.) relationships in order to more accurately categorize the ensemble and predictor periods.

## 6.2 Improving Short-term Forecasts using distributed hydrological model DA

Streamflow forecasting from a few days to approximately ten days is very sensitive to initial conditions, or 'boundaries' of the system, i.e., 'how good we can be at the starting point' (DeChant & Moradkhani, 2011; Li et al., 2009). Thus, issuing a forecast with a minimum error

or having the most accurate possible initial conditions is requested by streamflow forecasters since it improves forecast skill (Andreadis & Lettenmaier, 2006)

DA is a method that can improve the accuracy of initial conditions used in streamflow forecasting. It is a means to achieve "seamless" and "accurate" forecasts. Seamless in the sense that it integrates past information into a more coherent initial condition, and accurate because less error in the initial condition leads to more skillful streamflow forecasts (Clark et al., 2006; DeChant & Moradkhani, 2014).

In the second paper, it was noted that having a reasonable starting point is crucial for a seamless streamflow forecasting model. This means that the initial conditions should represent the real world, including factors such as soil moisture or any amount of water at different levels of the ground, which directly affect the performance of the forecasting model (Jadidoleslam et al., 2021; Prakash & Mishra, 2022). One method for achieving this is through EnKF (McMillan et al., 2013; Xie & Zhang, 2010). Applying the EnKF is not a straightforward task and comes with challenges, especially in tuning hyperparameters (Thiboult & Anctil, 2015). Adjusting hyperparameters is a delicate and extensive task (Clark et al., 2008). In addition, since performing DA can yield similar results for multiple hydrological models by correcting their errors, it might also limit the impact of using more or better hydrological models to a small extent (Bergeron et al., 2021).

In this study, we proposed a detailed procedure for seasonally tuning hyperparameters in a distributed operational hydrological model using the EnKF DA method. These hyperparameters include the uncertainties associated with inputs (e.g., temperature and precipitation), outputs (e.g., streamflow), and the choice of state variable. To address these uncertainties, we performed tests over a wide range of hyperparameters. We implemented the EnKF DA method for an operational distributed hydrological model and observed that certain hyperparameter sets showed good performance. Our evaluation revealed that the EnKF method

exhibited favorable performance and could be integrated into a probabilistic forecast framework.

Various factors were tested to optimize the application of the EnKF, and it was found that DA improved forecast skill in all seasons, with longer-lasting skill observed in the winter period. However, more difficulties were encountered in seasons with greater hydrological dynamics, such as spring (during melting runoff), or summer/fall (due to rainfall, evapotranspiration, and faster reaction vadose zone). Adjusting certain hyperparameters is required to apply the EnKF. These hyperparameters include the uncertainty of the hydrological model's inputs (e.g., temperature and precipitation), outputs (e.g., observed streamflow), and state variables. Changing these hyperparameters can affect the forecast skill.

The EnKF assimilation approach is ensemble-based and requires an ensemble of states to evolve over time from one time step to the next. In this study, a 50-member ensemble was chosen, which is more than the minimum of 40 members recommended by (Valdez et al., 2022) to balance stochastic error and computational costs. The perturbations or uncertainty of temperature were found to be the most dominant factor among the hydro-meteorological variables. In order to update the EnKF, high values of uncertainty for temperature data were necessary as lower values of uncertainty resulted in poor performance, as shown in Figure 4.3. This can be attributed to the effects of temperature on snow accumulation and melt (from the end of fall to the end of spring), as well as on the evapotranspiration rate (in summer). The uncertainty values for streamflow and precipitation were surprisingly lower, despite precipitation being typically harder to measure than temperature. However, even small changes in precipitation could have a larger impact on the model outputs, making temperature uncertainty more critical for providing the model with the flexibility it needs to match the observed streamflow.

The second paper demonstrated that significant levels of precipitation uncertainty can still yield good results in certain cases. This suggests that models may need to be pushed to their

limits to achieve optimal performance. However, we chose not to adopt these high-value results to avoid potential forecasting difficulties. Over-correcting the system may lead to penalties in forecasting accuracy. This highlights the need for supervision of the results. It is crucial to consider the realism of the results as adjusting the hyperparameters solely to improve the forecast skill can lead to unrealistic results, such as considering very high levels of uncertainty for precipitation. The results show that the DA had an impact for a maximum of six days before the model would start drifting again, which is similar to the catchment concentration time. The physical limitations of the modeling and forecast chain restrict the improvement of DA beyond lead-times longer than the concentration time, as shown by Chen et al. (2013). Patil and Ramsankaran (2017) demonstrated the importance of subsurface water in updating initial states, which is consistent with our findings as the vadose zone was the most sensitive state variable in the updating process. Updating the vadose zone is considered or proposed as the most important state variable, which underscores the significance of comprehending water in the unsaturated zone, subsurface waters, and groundwater (Drécourt et al., 2006; Noh et al., 2018).

## 6.3 Ability of neural network (LSTM) on short-term forecasting, while using proxies of DA

In the third part of our study, we compared the use of a RNN method with a hydrological model with DA. We then presented the results for streamflow modeling and forecasting. To simulate and forecast, we used a popular method called LSTM, which is a type of artificial neural network (Ghimire et al., 2021). We chose LSTM because its structure includes a "forget gate," which enables it to capture longer dependencies that are often neglected when using other types of artificial neural networks without this feature (Hunt et al., 2022; Mehedi et al., 2022; Schmidhuber & Hochreiter, 1997; Xu et al., 2020).

The use of observed streamflow as a predictor has been explored in previous literature (Cloke & Pappenberger, 2009). Historical streamflow has been used directly as a forecast scenario (e.g., Rajagopalan et al. (2010), as well as more complex methods (Bogner et al., 2016; Hopson

& Webster, 2010; Seo et al., 2006), and autoregressive methods (Amiri, 2015; Zhang et al., 2015). Artificial neural networks have also been investigated (Coulibaly et al., 2000; Machado et al., 2011). Recently, LSTM has received attention for its ability to simulate and predict streamflow. LSTM uses observed streamflow as a component to build a structure that functions like a process-based hydrological model.

Compared to DA, the new method shows lower dispersion of results at the beginning of the forecasts. The wider dispersion of DA results is a compensation for forecast uncertainty (Granata et al., 2022). Both methods were trained on historical data, and both results worsen with increasing lead-times, which is in line with the degradation of weather forecasts as well (Le et al., 2019).

The LSTM model used in this study was trained on historical observations of temperature, precipitation, and streamflow and used for forecasting lead-times ranging from one day to nine days. Since streamflow was included in the training phase, it was considered equivalent to the DA phase. One of the advantages of the LSTM model is that it does not require a separate DA step, making it more suitable for operational use. However, it does require separate training for each lead-time, which adds to the computational cost of forecasts. Like other methods, the LSTM model has difficulty in capturing extreme events and requires longer timeseries to "learn" more about unprecedented events (Zhu et al., 2020). Despite these limitations, the LSTM model showed relatively better performance compared to CEQUEAU-DA. Both methods showed a degradation of forecast skill as the lead-time increased, but while DA is limited to improving initial conditions, the LSTM model learns from hidden relationships between precipitation, temperature, and streamflow. This difference partially explains why the LSTM forecast results were better than DA forecast results. However, there are differences between the two methods, as the distribution of initial conditions is only available with DA. Additionally, while the LSTM model does not technically assimilate past data to improve initial states, it uses streamflow data along with temperature and precipitation. The lack of DA is an advantage of using the LSTM model, but the need for a trained model for each lead-time

is a limitation. The LSTM model's limitation in predicting extremes comes from not having enough data, which is why it requires a very long timeseries in which extreme events were recorded.

The LSTM results show less continuity compared to CEQUEAU-DA. This is because each lead-time is trained separately (see Figure 5.8). The discontinuity is also due to the fact that DA, by its nature, adds continuity to forecasts. However, it is worth noting that DA is not completely absent in LSTM. In fact, during the training of the LSTM model, observed streamflows are always used, so adaptation to real streamflow is always considered by the LSTM model. The equivalents to hydrological states in hydrological modeling are also updated.

Limitations in training LSTM models for specific lead-times can increase the amount of time and computations required. To address this limitation, one could include more variables to provide a more complete representation of the data. However, our results showed that this approach did not improve forecasts beyond a two-day lead-time. One possible explanation for this is the difference in sources between the historical data and the forecast data. Another option that we considered was to use forecasts from the previous day(s) as input, such as using the forecasts from the preceding three days for a four-day lead-time. However, this method also failed to improve the accuracy of forecasts beyond two days because errors in each forecast can propagate and accumulate when using multiple forecasts.

## 6.4 Seamless streamflow forecasting

This section discusses the efforts to improve the quality of streamflow forecasting in terms of accuracy and reliability for multiple lead-times. The study proposes various methods to achieve this objective, with three sub-objectives identified: improving long-term forecasts (i.e., multiple months lead-time) using ESP member filtering, improving short-term forecasts (i.e.,

up to 10 days) using distributed hydrological model DA, and evaluating the ability of data-driven models (LSTM) on short-term forecasting while using proxies of DA.

Improving streamflow forecasting is a challenging task due to the gap between short-term and long-term lead-times. The accuracy of meteorological forecasts decreases after a few days, and historical data are not always the best approach for long-term forecasting. In this regard, Sohrabi and Brissette (2021) and Sohrabi et al. (2021) proposed using a weather generator calibrated considering the correlation between large-scale indices to forecast streamflow seamlessly.

On the other hand, McInerney et al. (2020) and McInerney et al. (2022) suggested reflecting seasonal and inter-annual variability of precipitation and/or streamflow, modeling temporal characteristics, and using different ways of modeling errors to improve streamflow forecasting. In addition, Yuan et al. (2014) proposed combining forecasts from different sources to improve the accuracy and reliability of streamflow forecasts.

To achieve a comprehensive streamflow forecast, it is crucial to use DA to ensure that the starting point is as accurate as possible. This involves using a hydrological model that has been calibrated with meteorological forecasts and historical data. For short-term forecasts, recent information should be utilized to provide a more accurate representation of initial states and boundary conditions, which can be achieved through DA. However, care must be taken to ensure that initial state correction does not introduce new errors in the longer-term forecasts, as in Mai et al. (2020).

For subseasonal to seasonal forecasts, forecasts specifically issued for this period should be used. In contrast, for longer periods, historical data can serve as a baseline, but modifications should be made to account for current conditions. However, as climate conditions can change rapidly, even long-term forecasts may need to be updated after a period of time, and it is essential to address the need for the user. For instance, users may require information on large-

scale climate indices or the historical probability of having periods with specific amounts of water to filter scenarios and test their accuracy over hindcasts. To account for natural variability and improve streamflow forecasting accuracy, forecasters often use ensemble forecasting methods, which involve running multiple simulations with slightly different initial conditions or model parameters to generate a range of possible outcomes. This approach allows forecasters to assess the likelihood of different scenarios and provide more reliable forecasts in the face of inherent variability (Demaria et al., 2016; Fatichi et al., 2014; Seiller & Anctil, 2014; Wood et al., 2016; Zhao & Brissette, 2022).

Improving the accuracy and reliability of streamflow forecasts requires a combination of different approaches such as using calibrated hydrological models, DA, and combining forecasts from different sources while accounting for historical and current conditions. It is essential to consider the specific lead-time and user's needs to achieve a comprehensive streamflow forecast. Overall, improving streamflow forecasting is critical for effective water resources management and decision-making.

## 6.5 Limitations

The objective of this study was to improve both short-term and long-term streamflow forecasts. Short-term improvements were more attainable due to a good understanding of the weather over short time periods. However, this revealed limitations, such as the need to study multiple sites and sources of data to develop more robust forecasting methods. Long-term streamflow forecasting requires incorporation of the natural variability that exists in precipitation and temperature time series. This has been explored in previous studies, such as Kalra et al. (2013), Sagarika et al. (2014), Sohrabi et al. (2021), and Zhao and Brissette (2022). In addition, using LSTM models is an alternative since they have shown the ability to capture long-term temporal dependencies and can "learn" from the natural variability of temperature, precipitation, and other climate variables (Cheng et al., 2020; Herbert et al., 2021). Therefore, it is recommended to further investigate all lead-times using LSTM models, despite the challenges of learning

long-term dependencies (Hochreiter et al., 2001). This could be done with temperature and precipitation or with additional variables, such as soil moisture, cloud cover, and wind speed. It is important to note that using different variables sometimes requires different data sources, which highlights the need for bias correction and pre-processing as a first step.

On the other hand, the accuracy of long-term forecasts depends on the choice of the study region, which is another limitation of this study. The limited number of regions studied could be addressed in future studies by using the same setting as in chapter 4 but applying it to other regions to better evaluate the streamflow forecasting chain.

One limitation of DA for hydrological modeling is that it often relies on only one model, which may not capture the full complexity of the system. It could be beneficial to explore the use of multiple models to better capture uncertainty and improve accuracy, but this approach would require careful consideration of practical and computational challenges.

Another limitation is that while we have conducted both short-term and long-term forecasts, we have yet to issue seamless forecasts that cover a full range of lead-times. For example, it would be valuable to develop forecasts that span from 1-day to 365-day lead-times in a continuous manner, allowing for a more comprehensive understanding of the system dynamics and uncertainties over time.

# CONCLUSION

This study aimed to evaluate different methods for improving the accuracy of streamflow forecasting in Canadian catchments, which are snow-dominated catchments with spring floods. Depending on the lead-time of interest, there are generally two approaches. One approach is to use ensemble meteorological forecasts, such as precipitation and temperature, which perform well for up to ten days. In this study, an EnKF-based assimilation method was applied to improve the accuracy of forecasts for this ten-day period. The regulation of hyperparameters was addressed in seasonal settings. The results indicate that the seasonal DA scheme with fixed hyperparameters and the best state variables combination leads to an improvement in ensemble streamflow forecasts, especially in winter. However, the skill of the ensemble forecast decreases with longer lead-times. A sensitivity analysis was conducted to evaluate the influence of the hyperparameters and model state variables, and the results show strong seasonal sensitivity of the EnKF hyperparameters and model state variables. The study suggests that a detailed testing of all possible combinations could help forecasters better target the assimilation hyperparameters for their specific needs. The EnKF assimilation scheme has significant potential to improve streamflow forecasts for operational forecasting centers and water resources system managers.

The benchmark dataset was used to evaluate the performance of LSTM models for forecasting short-term streamflow. The LSTM model provided more skillful ensemble forecasts and displayed tighter spreads compared to CEQUEAU. It eliminated the need for a DA process and still provided high-quality forecasts for lead-times up to 9 days, with a KGE higher than 0.88. The study concludes that LSTM is a promising model for short-term streamflow forecasting, and it can provide better accuracy than more complex distributed hydrological models in this North American catchment. The results showed that LSTM outperformed the use of DA until a lead-time of 5 days.

The other part of the study aimed to improve long-term streamflow forecasting using historical meteorological data as possible forecast scenarios. These scenarios needed to be processed to obtain better forecasts by considering the characteristics of the period before the forecast date. Therefore, a GA-based method was used to quantify the maximum potential improvement and provide a means of comparison between various methods. The study proposed a method to differentiate desirable and undesirable members by correlating hydroclimatological indices with the theoretically optimal ESP member properties. However, the selected variables were not found to be good predictors of this relationship. The findings could help improve ESP forecasting by providing a more consistent baseline for comparison and determining the potential of streamflow forecasting using historical data. Additionally, indexing historical scenarios could help select the optimal set and use it as a proxy to filter scenarios, thereby improving the data pre-processing.

The study showed that various methods could be used to improve streamflow forecasting accuracy. The ESP member selection approach was useful for long-term forecasting, while the EnKF and LSTM models were effective for short-term forecasting. The study emphasized the importance of conducting detailed sensitivity analyses to identify the best-performing implementation of the EnKF model, its hyperparameters, and state variable selection for the specific needs of operational forecasters and water resources system managers.

## RECOMMENDATIONS

The purpose of this study is to increase the precision of short-term and long-term streamflow forecasts. Short-term improvements are more achievable due to a comprehensive grasp of local weather patterns. However, limitations such as the need to gather data from various locations and sources to develop stronger methods have been identified. The accuracy of long-term forecasts heavily depends on the climate and meteorological characteristics of the area in question and thus requires further examination, including the incorporation of additional climatic variables. These limitations underscore the importance of further research to incorporate data from multiple sources to establish more robust forecasting methods.

In this study, we have demonstrated that LSTM models are as effective, if not more so, than hydrological models based on physical processes. Furthermore, LSTM models can capture the temporal relationship between precipitation and streamflow. There is significant potential for enhancing LSTM models through techniques such as bias correction, conditioning, and weighting during pre-processing, as well as applying LSTMs at different temporal resolutions or improving their structures (Li et al., 2021a, 2021b). Further research is required to assess the utility of LSTM models in streamflow forecasting, including variations in model structure, input parameters, lead-times, and training in various regions. An objective of this study should be to compare and enhance the performance of LSTM models in comparison to the commonly used physical process-based hydrological models for streamflow forecasting (Arsenault et al., 2023; Feng et al., 2020; Liu et al., 2022).

The potential of LSTMs to capture the temporal dependency of precipitation and/or streamflow can be evaluated by using them to forecast streamflows over extended lead-times. Further study is necessary in different settings. In order to improve the estimation of long-term streamflows, it would be interesting to explore the capability of LSTMs to predict existing climate signals and streamflow events and then apply that knowledge to forecasting scenarios.

It is recommended to utilize other sources of data in different ways for future studies. For instance, one may consider training LSTM on other basins, either with or without similarities. Also, the use of remote sensing data and other sources of information, such as social media and citizen science, could also be explored as potential inputs for Artificial Intelligence (AI) and Machine Learning (ML) models. By combining data from various sources, it may be possible to develop more accurate and robust forecasting models, particularly for areas with limited ground-based monitoring networks. However, further research is needed to assess the quality and usefulness of these data sources for streamflow forecasting and to develop methods for integrating them with traditional data sources.

The coupling of drought studies and streamflow predictions is crucial for meeting the need for long-term streamflow forecasting in dry regions. Accurate forecasting is important for preventing the human costs of water scarcity, and advance notice can provide a clearer picture of water resources, which is especially valuable in regions facing both water scarcity and political conflicts.

# APPENDIX I CEQUEAU MODEL

This schematic and parameter sets were adopted and translated from the documentation provided by Rio Tinto (*Développements du nouveau CEQUEAU*).

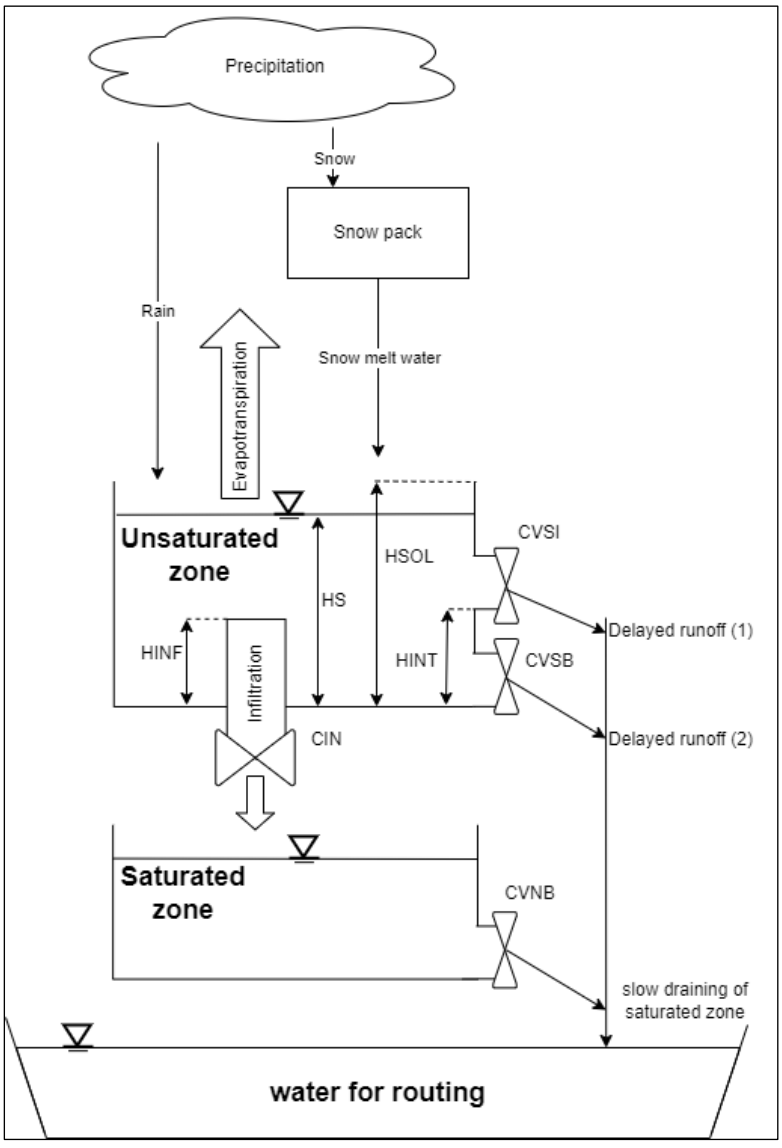


Figure A I Production diagram of the Cequeau model.

Table A I CEQUEAU parameters

| Parameters | |
|---|---|
| Snow melt parameters (CemaNeige) | |
| Kf | Melting rate (mm/°C) |
| Tf | Melting temperature (°C) |
| CTg | Thermal state weighting coefficient (%) |
| Gseuil | Minimum snow stock (mm) |
| Vmin | Minimum melting of potential melting (%) |
| strne | Snow-to-water transformation threshold (°C) |
| Evapotranspiration parameters (Oudin) | |
| K1 | Scaling factor for adjusting the ETP volume |
| K2 | Temperature threshold (°C) |
| XLA | Average latitude of the partial tile (in degree-minute) |
| Production parameters: water in unsaturated zone and water in saturated zone | |
| Cin | Infiltration coefficient from the unsaturated zone to the saturated zone |
| Cvnb | Low drainage coefficient of the saturated zone |
| Cvsi | Intermediate drainage coefficient of the unsaturated zone |
| Cvsb | Low drainage coefficient of the unsaturated zone |
| Hinf | Infiltration height from from the unsaturated zone to the saturated zone |
| Hint | Intermediate drainage height of the unsaturated zone |
| Hsol | Height of the unsaturated zone (in mm) |
| Xinfma | Maximum infiltration per day |
| Vidintmax | Maximum threshold for intermediate drainage of the unsaturated zone |
| Transfer parameters | |
| HUDebit | Unit hydrograph for flow transfer |
| HUProd | Unit hydrograph for flow attenuation from its outlet to the downstream basin outlet |

## LIST OF BIBLIOGRAPHICAL REFERENCES


Abaza, M., Anctil, F., Fortin, V., & Perreault, L. (2017). Hydrological evaluation of the canadian meteorological ensemble reforecast product. *Atmosphere-Ocean*, *55*(3), 195-211. https://doi.org/10.1080/07055900.2017.1341384

Abbasnezhadi, K., Rousseau, A. N., Foulon, É., & Savary, S. (2021). Verification of Regional Deterministic Precipitation Analysis Products Using Snow Data Assimilation for Application in Meteorological Network Assessment in Sparsely Gauged Nordic Basins. *Journal of Hydrometeorology*, *22*(4), 859-876. https://doi.org/10.1175/JHM-D-20-0106.1

Abbaszadeh, P., Moradkhani, H., & Yan, H. (2018). Enhancing hydrologic data assimilation by evolutionary particle filter and Markov chain Monte Carlo. *Advances in Water Resources*, *111*, 192-204. https://doi.org/https://doi.org/10.1016/j.advwatres.2017.11.011

Ajami, N. K., Duan, Q. Y., & Sorooshian, S. (2007). An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction. *Water Resources Research*, *43*(1). https://doi.org/10.1029/2005wr004745

Alemu, E. T., Palmer, R. N., Polebitski, A., & Meaker, B. (2011). Decision Support System for Optimizing Reservoir Operations Using Ensemble Streamflow Predictions. *Journal of Water Resources Planning and Management*, *137*(1), 72-82. https://doi.org/10.1061/(Asce)Wr.1943-5452.0000088

Alfieri, L., Pappenberger, F., Wetterhall, F., Haiden, T., Richardson, D., & Salamon, P. (2014). Evaluation of ensemble streamflow predictions in Europe. *Journal of Hydrology*, *517*, 913-922. https://doi.org/10.1016/j.jhydrol.2014.06.035

Amiri, E. (2015). Forecasting daily river flows using nonlinear time series models. *Journal of Hydrology*, *527*, 1054-1072. https://doi.org/10.1016/j.jhydrol.2015.05.048

Anctil, F., Perrin, C., & Andréassian, V. (2003). ANN OUTPUT UPDATING OF LUMPED CONCEPTUAL RAINFALL/RUNOFF FORECASTING MODELS 1. *JAWRA Journal of the American Water Resources Association*, *39*(5), 1269-1279. https://doi.org/10.1111/j.1752-1688.2003.tb03708.x

Andreadis, K. M., & Lettenmaier, D. P. (2006). Assimilating remotely sensed snow observations into a macroscale hydrology model. *Advances in Water Resources*, *29*(6), 872-886. https://doi.org/10.1016/j.advwatres.2005.08.004

Anghileri, D., Voisin, N., Castelletti, A., Pianosi, F., Nijssen, B., & Lettenmaier, D. P. (2016). Value of long-term streamflow forecasts to reservoir operations for water supply in snow-dominated river catchments. *Water Resources Research*, *52*(6), 4209-4225. https://doi.org/https://doi.org/10.1002/2015WR017864

Araghinejad, S., Burn, D. H., & Karamouz, M. (2006). Long-lead probabilistic forecasting of streamflow using ocean-atmospheric and hydrological predictors. *Water Resources Research*, *42*(3). https://doi.org/10.1029/2004WR003853

Arnal, L., Cloke, H. L., Stephens, E., Wetterhall, F., Prudhomme, C., Neumann, J., Krzeminski, B., & Pappenberger, F. (2018). Skilful seasonal forecasts of streamflow over Europe? *Hydrology and Earth System Sciences*, *22*(4), 2057-2072. https://doi.org/https://doi.org/10.5194/hess-22-2057-2018

Arsenault, R., & Brissette, F. (2016). Multi-model averaging for continuous streamflow prediction in ungauged basins. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, *61*(13), 2443-2454. https://doi.org/10.1080/02626667.2015.1117088

Arsenault, R., Brissette, F., & Martel, J.-L. (2018). The hazards of split-sample validation in hydrological model calibration. *Journal of Hydrology*, *566*, 346-362. https://doi.org/10.1016/j.jhydrol.2018.09.027

Arsenault, R., & Côté, P. (2019). Analysis of the effects of biases in ensemble streamflow prediction (ESP) forecasts on electricity production in hydropower reservoir management. *Hydrology and Earth System Sciences*, *23*(6), 2735-2750. https://doi.org/10.5194/hess-23-2735-2019

Arsenault, R., Martel, J.-L., Brunet, F., Brissette, F., & Mai, J. (2023). Continuous streamflow prediction in ungauged basins: long short-term memory neural networks clearly outperform traditional hydrological models. *Hydrology and Earth System Sciences*, *27*(1), 139-157. https://doi.org/10.5194/hess-27-139-2023

Baker, S. A., Rajagopalan, B., & Wood, A. W. (2021). Enhancing Ensemble Seasonal Streamflow Forecasts in the Upper Colorado River Basin Using Multi-Model Climate Forecasts. *Journal of the American Water Resources Association*, *57*(6), 906-922. https://doi.org/10.1111/1752-1688.12960

Bao, H.-J., Zhao, L.-N., He, Y., Li, Z.-J., Wetterhall, F., Cloke, H., Pappenberger, F., & Manful, D. (2011). Coupling ensemble weather predictions based on TIGGE database

with Grid-Xinanjiang model for flood forecast. *Advances in Geosciences*, *29*, 61-67. https://doi.org/10.5194/adgeo-29-61-2011

Beckers, J. V., Weerts, A. H., Tijdeman, E., & Welles, E. (2016). ENSO-conditioned weather resampling method for seasonal ensemble streamflow prediction. *Hydrology and Earth System Sciences*, *20*(8), 3277-3287. https://doi.org/10.5194/hess-20-3277-2016

Bellier, J., Bontron, G., & Zin, I. (2017). Using Meteorological Analogues for Reordering Postprocessed Precipitation Ensembles in Hydrological Forecasting. *Water Resources Research*, *53*(12), 10085-10107. https://doi.org/10.1002/2017wr021245

Bellier, J., Zin, I., & Bontron, G. (2018). Generating Coherent Ensemble Forecasts After Hydrological Postprocessing: Adaptations of ECC-Based Methods. *Water Resources Research*, *54*(8), 5741-5762. https://doi.org/10.1029/2018wr022601

Bellier, J., Zin, I., Siblot, S., & Bontron, G. (2016). Probabilistic flood forecasting on the Rhone River: evaluation with ensemble and analogue-based precipitation forecasts. E3S Web of Conferences, https://doi.org/10.1051/e3sconf/20160718011

Bennett, J. C., Wang, Q. J., Pokhrel, P., & Robertson, D. E. (2014). The challenge of forecasting high streamflows 1-3 months in advance with lagged climate indices in southeast Australia. *Natural Hazards and Earth System Sciences*, *14*(2), 219-233. https://doi.org/10.5194/nhess-14-219-2014

Bergeron, J., Leconte, R., Trudel, M., & Farhoodi, S. (2021). On the Choice of Metric to Calibrate Time-Invariant Ensemble Kalman Filter Hyper-Parameters for Discharge Data Assimilation and Its Impact on Discharge Forecast Modelling. *Hydrology*, *8*(1), 36. https://doi.org/https://doi.org/10.3390/hydrology8010036

Bogner, K., Cloke, H., Pappenberger, F., De Roo, A., & Thielen, J. (2012). Improving the evaluation of hydrological multi-model forecast performance in the Upper Danube Catchment. *International journal of river basin management*, *10*(1), 1-12. https://doi.org/10.1080/15715124.2011.625359

Bogner, K., Liechti, K., & Zappa, M. (2016). Post-Processing of Stream Flows in Switzerland with an Emphasis on Low Flows and Floods. *Water*, *8*(4), 115. https://doi.org/10.3390/w8040115

Bojariu, R., & Gimeno, L. (2003). The role of snow cover fluctuations in multiannual NAO persistence. *Geophysical Research Letters*, *30*(4). https://doi.org/10.1029/2002GL015651

Boucher, M.-A., Anctil, F., Perreault, L., & Tremblay, D. (2011). A comparison between ensemble and deterministic hydrological forecasts in an operational context. *Advances in Geosciences*, *29*, 85-94. https://doi.org/10.5194/adgeo-29-85-2011

Boucher, M.-A., Tremblay, D., Delorme, L., Perreault, L., & Anctil, F. (2012). Hydro-economic assessment of hydrological forecasting systems. *Journal of Hydrology*, *416*, 133-144. https://doi.org/10.1016/j.jhydrol.2011.11.042

Bourgin, F., Ramos, M. H., Thirel, G., & Andreassian, V. (2014). Investigating the interactions between data assimilation and post-processing in hydrological ensemble forecasting. *Journal of Hydrology*, *519*, 2775-2784. https://doi.org/10.1016/j.jhydrol.2014.07.054

Bradley, A. A., Habib, M., & Schwartz, S. S. (2015). Climate index weighting of ensemble streamflow forecasts using a simple B ayesian approach. *Water Resources Research*, *51*(9), 7382-7400. https://doi.org/10.1002/2014WR016811

Brajard, J., Carrassi, A., Bocquet, M., & Bertino, L. (2020). Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the Lorenz 96 model. *Journal of Computational Science*, *44*, 101171. https://doi.org/10.1016/j.jocs.2020.101171

Brochero, D., Anctil, F., & Gagne, C. (2011). Simplifying a hydrological ensemble prediction system with a backward greedy selection of members - Part 1: Optimization criteria. *Hydrology and Earth System Sciences*, *15*(11), 3307-3325. https://doi.org/10.5194/hess-15-3307-2011

Buizza, R., Du, J., Toth, Z., & Hou, D. (2018). Major operational ensemble prediction systems (EPS) and the future of EPS. *Handbook of hydrometeorological ensemble forecasting*, 151-193. https://doi.org/10.1007/978-3-642-40457-3_14-1

Caillouet, L., Celie, S., Vannier, O., Bontron, G., & Legrand, S. (2022). Operational hydrometeorological forecasting on the Rhone River in France: moving toward a seamless probabilistic approach. *Lhb-Hydroscience Journal*, *108*(1), 2061312. https://doi.org/10.1080/27678490.2022.2061312

Cassagnole, M., Ramos, M. H., Zalachori, I., Thirel, G., Garcon, R., Gailhard, J., & Ouillon, T. (2021). Impact of the quality of hydrological forecasts on the management and revenue of hydroelectric reservoirs - a conceptual approach. *Hydrology and Earth System Sciences*, *25*(2), 1033-1052. https://doi.org/10.5194/hess-25-1033-2021

Ceola, S., Montanari, A., & Koutsoyiannis, D. (2014). Toward a theoretical framework for integrated modeling of hydrological change. *Wiley Interdisciplinary Reviews-Water*, *1*(5), 427-438. https://doi.org/10.1002/wat2.1038

Chandimala, J., & Zubair, L. (2007). Predictability of stream flow and rainfall based on ENSO for water resources management in Sri Lanka. *Journal of Hydrology*, *335*(3-4), 303-312. https://doi.org/10.1016/j.jhydrol.2006.11.024

Chen, H., Yang, D. W., Hong, Y., Gourley, J. J., & Zhang, Y. (2013). Hydrological data assimilation with the Ensemble Square-Root-Filter: Use of streamflow observations to update model states for real-time flash flood forecasting. *Advances in Water Resources*, *59*, 209-220. https://doi.org/10.1016/j.advwatres.2013.06.010

Chen, Y., Li, J., & Xu, H. (2016). Improving flood forecasting capability of physically based distributed hydrological models by parameter optimization. *Hydrology and Earth System Sciences*, *20*(1), 375-392. https://doi.org/10.5194/hess-20-375-2016

Cheng, M., Fang, F., Kinouchi, T., Navon, I. M., & Pain, C. C. (2020). Long lead-time daily and monthly streamflow forecasting using machine learning methods. *Journal of Hydrology*, *590*, 125376. https://doi.org/10.1016/j.jhydrol.2020.125376

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*. https://doi.org/10.3115/v1/D14-1179

Clark, M. P., & Hay, L. E. (2004). Use of medium-range numerical weather prediction model output to produce forecasts of streamflow. *Journal of Hydrometeorology*, *5*(1), 15-32. https://doi.org/10.1175/1525-7541(2004)005%3C0015:UOMNWP%3E2.0.CO;2

Clark, M. P., Rupp, D. E., Woods, R. A., Zheng, X., Ibbitt, R. P., Slater, A. G., Schmidt, J., & Uddstrom, M. J. (2008). Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. *Advances in Water Resources*, *31*(10), 1309-1324. https://doi.org/10.1016/j.advwatres.2008.06.005

Clark, M. P., Slater, A. G., Barrett, A. P., Hay, L. E., McCabe, G. J., Rajagopalan, B., & Leavesley, G. H. (2006). Assimilation of snow covered area information into hydrologic and land-surface models. *Advances in Water Resources*, *29*(8), 1209-1221. https://doi.org/10.1016/j.advwatres.2005.10.001

Cloke, H. L., & Pappenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, *375*(3-4), 613-626. https://doi.org/10.1016/j.jhydrol.2009.06.005

Collischonn, W., Haas, R., Andreolli, I., & Tucci, C. E. M. (2005). Forecasting River Uruguay flow using rainfall forecasts from a regional weather-prediction model. *Journal of Hydrology*, *305*(1-4), 87-98. https://doi.org/10.1016/j.jhydrol.2004.08.028

Coulibaly, P., Anctil, F., & Bobee, B. (2000). Daily reservoir inflow forecasting using artificial neural networks with stopped training approach. *Journal of Hydrology*, *230*(3-4), 244-257. https://doi.org/10.1016/S0022-1694(00)00214-6

Crochemore, L., Ramos, M.-H., Pappenberger, F., & Perrin, C. (2017). Seasonal streamflow forecasting by conditioning climatology with precipitation indices. *Hydrology and Earth System Sciences*, *21*(3), 1573-1591. https://doi.org/10.5194/hess-21-1573-2017

Cuo, L., Beyene, T. K., Voisin, N., Su, F. G., Lettenmaier, D. P., Alberti, M., & Richey, J. E. (2011). Effects of mid-twenty-first century climate and land cover change on the hydrology of the Puget Sound basin, Washington. *Hydrological Processes*, *25*(11), 1729-1753. https://doi.org/10.1002/hyp.7932

Damavandi, H. G., Shah, R., Stampoulis, D., Wei, Y., Boscovic, D., & Sabo, J. (2019). Accurate prediction of streamflow using long short-term memory network: a case study in the Brazos River Basin in Texas. *International Journal of Environmental Science and Development*, *10*(10), 294-300. https://doi.org/10.18178/ijesd.2019.10.10.1190

Day, G. N. (1985). Extended streamflow forecasting using NWSRFS. *Journal of Water Resources Planning and Management*, *111*(2), 157-170. https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157)

DeChant, C. M., & Moradkhani, H. (2011). Improving the characterization of initial condition for ensemble streamflow prediction using data assimilation. *Hydrology and Earth System Sciences*, *15*(11), 3399-3410. https://doi.org/10.5194/hess-15-3399-2011

DeChant, C. M., & Moradkhani, H. (2014). Toward a reliable prediction of seasonal forecast uncertainty: Addressing model and initial condition uncertainty with ensemble data assimilation and Sequential Bayesian Combination. *Journal of Hydrology*, *519*, 2967-2977. https://doi.org/10.1016/j.jhydrol.2014.05.045

Demargne, J., Mullusky, M., Werner, K., Adams, T., Lindsey, S., Schwein, N., Marosi, W., & Welles, E. (2009). Application of forecast verification science to operational river forecasting in the US National Weather Service. *Bulletin of the American Meteorological Society*, *90*(6), 779-784. https://doi.org/10.1175/2008BAMS2619.1

Demaria, E. M., Palmer, R. N., & Roundy, J. K. (2016). Regional climate change projections of streamflow characteristics in the Northeast and Midwest US. *Journal of Hydrology: Regional Studies*, *5*, 309-323. https://doi.org/10.1016/j.ejrh.2015.11.007

Dettinger, M. D., Cayan, D. R., Diaz, H. F., & Meko, D. M. (1998). North–south precipitation patterns in western North America on interannual-to-decadal timescales. *Journal of Climate*, *11*(12), 3095-3111. https://doi.org/10.1175/1520-0442(1998)011%3C3095:NSPPIW%3E2.0.CO;2

Devineni, N., Sankarasubramanian, A., & Ghosh, S. (2008). Multimodel ensembles of streamflow forecasts: Role of predictor state in developing optimal combinations. *Water Resources Research*, *44*(9). https://doi.org/10.1029/2006WR005855

Dion, P., Martel, J.-L., & Arsenault, R. (2021). Hydrological ensemble forecasting using a multi-model framework. *Journal of Hydrology*, *600*, 126537. https://doi.org/10.1016/j.jhydrol.2021.126537

Doblas-Reyes, F. J., García-Serrano, J., Lienert, F., Biescas, A. P., & Rodrigues, L. R. (2013). Seasonal climate predictability and forecasting: status and prospects. *Wiley Interdisciplinary Reviews: Climate Change*, *4*(4), 245-268. https://doi.org/10.1002/wcc.217

Donegan, S., Murphy, C., Harrigan, S., Broderick, C., Foran Quinn, D., Golian, S., Knight, J., Matthews, T., Prudhomme, C., & Scaife, A. A. (2021). Conditioning ensemble streamflow prediction with the North Atlantic Oscillation improves skill at longer lead times. *Hydrology and Earth System Sciences*, *25*(7), 4159-4183. https://doi.org/10.5194/hess-25-4159-2021

Drécourt, J.-P., Madsen, H., & Rosbjerg, D. (2006). Calibration framework for a Kalman filter applied to a groundwater model. *Advances in Water Resources*, *29*(5), 719-734. https://doi.org/10.1016/j.advwatres.2005.07.007

Duan, Q., Ajami, N. K., Gao, X., & Sorooshian, S. (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Advances in Water Resources*, *30*(5), 1371-1386. https://doi.org/10.1016/j.advwatres.2006.11.014

Ehsan, M. A., Tippett, M. K., Robertson, A. W., Almazroui, M., Ismail, M., Dinku, T., Acharya, N., Siebert, A., Ahmed, J. S., & Teshome, A. (2021). Seasonal predictability of Ethiopian Kiremt rainfall and forecast skill of ECMWF's SEAS5 model. *Climate Dynamics*, *57*(11), 3075-3091. https://doi.org/10.1007/s00382-021-05855-0

Emerton, R., Zsoter, E., Arnal, L., Cloke, H. L., Muraro, D., Prudhomme, C., Stephens, E. M., Salamon, P., & Pappenberger, F. (2018). Developing a global operational seasonal hydro-meteorological forecasting system: GloFAS-Seasonal v1. 0. *Geoscientific Model Development*, *11*(8), 3327-3346. https://doi.org/10.5194/gmd-11-3327-2018

Evensen, G. (1994). Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *Journal of Geophysical Research: Oceans*, *99*(C5), 10143-10162. https://doi.org/10.1029/94JC00572

Fan, F. M., Schwanenberg, D., Alvarado, R., Assis dos Reis, A., Collischonn, W., & Naumman, S. (2016). Performance of deterministic and probabilistic hydrological forecasts for the short-term optimization of a tropical hydropower reservoir. *Water Resources Management*, *30*(10), 3609-3625. https://doi.org/10.1007/s11269-016-1377-8

Fang, K., & Shen, C. (2020). Near-real-time forecast of satellite-based soil moisture using long short-term memory with an adaptive data integration kernel. *Journal of Hydrometeorology*, *21*(3), 399-413. https://doi.org/10.1175/JHM-D-19-0169.1

Fatichi, S., Rimkus, S., Burlando, P., & Bordoy, R. (2014). Does internal climate variability overwhelm climate change signals in streamflow? The upper Po and Rhone basin case studies. *Science of the Total Environment*, *493*, 1171-1182. https://doi.org/10.1016/j.scitotenv.2013.12.014

Feng, D., Fang, K., & Shen, C. (2020). Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *Water Resources Research*, *56*(9), e2019WR026793. https://doi.org/10.1029/2019WR026793

Feng, Q. Y., Vasile, R., Segond, M., Gozolchiani, A., Wang, Y., Abel, M., Havlin, S., Bunde, A., & Dijkstra, H. A. (2016). ClimateLearn: A machine-learning approach for climate prediction using network measures. *Geoscientific Model Development Discussions*, 1-18. https://doi.org/10.5194/gmd-2015-273

Ferro, C. A., Richardson, D. S., & Weigel, A. P. (2008). On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, *15*(1), 19-24. https://doi.org/10.1002/met.45

Ficchì, A., Raso, L., Dorchies, D., Pianosi, F., Malaterre, P.-O., Van Overloop, P.-J., & Jay-Allemand, M. (2016). Optimal operation of the multireservoir system in the seine river basin using deterministic and ensemble forecasts. *Journal of Water Resources Planning and Management*, *142*(1), 05015005. https://doi.org/10.1061/(ASCE)WR.1943-5452.0000571

Friederichs, P., & Hense, A. (2007). Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly weather review*, *135*(6), 2365-2378. https://doi.org/10.1175/MWR3403.1

Fundel, F., Jörg-Hess, S., & Zappa, M. (2013). Monthly hydrometeorological ensemble prediction of streamflow droughts and corresponding drought indices. *Hydrology and Earth System Sciences*, *17*(1), 395-407. https://doi.org/10.5194/hess-17-395-2013

Ghimire, S., Deo, R. C., Wang, H., Al-Musaylh, M. S., Casillas-Pérez, D., & Salcedo-Sanz, S. (2022). Stacked LSTM Sequence-to-Sequence Autoencoder with Feature Selection for Daily Solar Radiation Prediction: A Review and New Modeling Results. *Energies*, *15*(3), 1061. https://www.mdpi.com/1996-1073/15/3/1061

Ghimire, S., Yaseen, Z. M., Farooque, A. A., Deo, R. C., Zhang, J., & Tao, X. (2021). Streamflow prediction using an integrated methodology based on convolutional neural network and long short-term memory networks. *Scientific Reports*, *11*(1). https://doi.org/10.1038/s41598-021-96751-4

Girihagama, L., Naveed Khaliq, M., Lamontagne, P., Perdikaris, J., Roy, R., Sushama, L., & Elshorbagy, A. (2022). Streamflow modelling and forecasting for Canadian watersheds using LSTM networks with attention mechanism. *Neural Computing and Applications*, 1-21. https://doi.org/10.1007/s00521-022-07523-8

Gneiting, T., Balabdaoui, F., & Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(2), 243-268. https://doi.org/10.1111/j.1467-9868.2007.00587.x

Granata, F., Di Nunno, F., & de Marinis, G. (2022). Stacked machine learning algorithms and bidirectional long short-term memory networks for multi-step ahead streamflow forecasting: A comparative study. *Journal of Hydrology*, *613*, 128431. https://doi.org/10.1016/j.jhydrol.2022.128431

Grantz, K., Rajagopalan, B., Clark, M., & Zagona, E. (2005). A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts. *Water Resources Research*, *41*(10). https://doi.org/10.1029/2004WR003467

Gupta, H. V., Kling, H., Yilmaz, K. K., & 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. https://doi.org/10.1016/j.jhydrol.2009.08.003

Gutiérrez, F., & Dracup, J. (2001). An analysis of the feasibility of long-range streamflow forecasting for Colombia using El Nino–Southern Oscillation indicators. *Journal of Hydrology*, *246*(1-4), 181-196. https://doi.org/10.1016/S0022-1694(01)00373-0

Habets, F., LeMoigne, P., & Noilhan, J. (2004). On the utility of operational precipitation forecasts to served as input for streamflow forecasting. *Journal of Hydrology*, *293*(1-4), 270-288. https://doi.org/10.1016/j.jhydrol.2004.02.004

Hamill, T. M. (2018). Practical aspects of statistical postprocessing. In *Statistical postprocessing of ensemble forecasts* (pp. 187-217). Elsevier. https://doi.org/10.1175/MWR3237.1

Hamlet, A. F., & Lettenmaier, D. P. (1999). Columbia River streamflow forecasting based on ENSO and PDO climate signals. *Journal of Water Resources Planning and Management*, *125*(6), 333-341. https://doi.org/10.1061/(ASCE)0733-9496(1999)125:6(333)

Hao, Z., Singh, V. P., & Xia, Y. (2018). Seasonal drought prediction: advances, challenges, and future prospects. *Reviews of Geophysics*, *56*(1), 108-141. https://doi.org/10.1002/2016RG000549

Hapuarachchi, H. A. P., Bari, M. A., Kabir, A., Hasan, M. M., Woldemeskel, F. M., Gamage, N., Sunter, P. D., Zhang, X. S., Robertson, D. E., & Bennett, J. C. (2022). Development of a national 7-day ensemble streamflow forecasting service for Australia. *Hydrology and Earth System Sciences*, *26*(18), 4801-4821. https://doi.org/10.5194/hess-26-4801-2022

Harou, J. J., Pulido-Velazquez, M., Rosenberg, D. E., Medellín-Azuara, J., Lund, J. R., & Howitt, R. E. (2009). Hydro-economic models: Concepts, design, applications, and future prospects. *Journal of Hydrology*, *375*(3-4), 627-643. https://doi.org/10.1016/j.jhydrol.2009.06.037

Harrigan, S., Prudhomme, C., Parry, S., Smith, K., & Tanguy, M. (2018). Benchmarking ensemble streamflow prediction skill in the UK. *Hydrology and Earth System Sciences*, *22*(3), 2023-2039. https://doi.org/10.5194/hess-22-2023-2018

Hashino, T., Bradley, A., & Schwartz, S. (2007). Evaluation of bias-correction methods for ensemble streamflow volume forecasts. *Hydrology and Earth System Sciences*, *11*(2), 939-950. https://doi.org/10.5194/hess-11-939-2007

He, M., Whitin, B., Hartman, R., Henkel, A., Fickenschers, P., Staggs, S., Morin, A., Imgarten, M., Haynes, A., & Russo, M. (2016). Verification of ensemble water supply forecasts for Sierra Nevada watersheds. *Hydrology*, *3*(4), 35. https://doi.org/10.3390/hydrology3040035

Hemri, S., & Klein, B. (2017). Analog-based postprocessing of navigation-related hydrological ensemble forecasts. *Water Resources Research*, *53*(11), 9059-9077. https://doi.org/10.1002/2017WR020684

Hendricks Franssen, H.-J., & Kinzelbach, W. (2008). Real-time groundwater flow modeling with the ensemble Kalman filter: Joint estimation of states and parameters and the filter inbreeding problem. *Water Resources Research*, *44*(9). https://doi.org/10.1029/2007WR006505

Herbert, Z. C., Asghar, Z., & Oroza, C. A. (2021). Long-term reservoir inflow forecasts: enhanced water supply and inflow volume accuracy using deep learning. *Journal of Hydrology*, *601*, 126676. https://doi.org/10.1016/j.jhydrol.2021.126676

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559-570. https://doi.org/10.1175/1520-0434(2000)015%3C0559:DOTCRP%3E2.0.CO;2

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., & Schepers, D. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, *146*(730), 1999-2049. https://doi.org/10.1002/qj.3803

Hirsch, R. M., Schaake, J. C., & Sheer, D. P. (1977). *Assessment of the current Occoquan water-supply situation, Fairfax County, Virginia* (2331-1258). https://doi.org/10.3133/ofr77811

Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In: A field guide to dynamical recurrent neural networks. IEEE Press In. https://doi.org/10.1162/neco.1997.9.8.1735

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hopson, T. M., & Webster, P. J. (2010). A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: Forecasting severe floods of 2003–07. *Journal of Hydrometeorology*, *11*(3), 618-641. https://doi.org/10.1175/2009JHM1006.1

Hossain, I., Esha, R., & Alam Imteaz, M. (2018). An attempt to use non-linear regression modelling technique in long-term seasonal rainfall forecasting for australian capital territory. *Geosciences*, *8*(8), 282. https://doi.org/10.3390/geosciences8080282

Hossain, I., Rasel, H., Imteaz, M. A., & Mekanik, F. (2020). Long-term seasonal rainfall forecasting using linear and non-linear modelling approaches: a case study for Western

Australia. *Meteorology and Atmospheric Physics*, *132*(1), 131-141. https://doi.org/10.1007/s00703-019-00679-4

Hu, C., Wu, Q., Li, H., Jian, S., Li, N., & Lou, Z. (2018). Deep Learning with a Long Short-Term Memory Networks Approach for Rainfall-Runoff Simulation. *Water*, *10*(11), 1543. https://doi.org/10.3390/w10111543

Hudson, D., Alves, O., Hendon, H. H., Lim, E.-P., Liu, G., Luo, J.-J., MacLachlan, C., Marshall, A. G., Shi, L., & Wang, G. (2017). ACCESS-S1 The new Bureau of Meteorology multi-week to seasonal prediction system. *Journal of Southern Hemisphere Earth Systems Science*, *67*(3), 132-159. https://doi.org/10.1071/ES17009_CO

Hunt, K. M. R., Matthews, G. R., Pappenberger, F., & Prudhomme, C. (2022). *Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States*. Copernicus GmbH. https://dx.doi.org/10.5194/hess-2022-53

Hurrell, J. W., Kushnir, Y., Ottersen, G., & Visbeck, M. (2008). The North Atlantic Oscillation: climatic significance and environmental impact. AGU Fall Meeting Abstracts,

Hutchinson, M. F., McKenney, D. W., Lawrence, K., Pedlar, J. H., Hopkinson, R. F., Milewska, E., & Papadopol, P. (2009). Development and testing of Canada-wide interpolated spatial models of daily minimum–maximum temperature and precipitation for 1961–2003. *Journal of Applied Meteorology and Climatology*, *48*(4), 725-741. https://doi.org/10.1175/2008JAMC1979.1

Jadidoleslam, N., Mantilla, R., & Krajewski, W. F. (2021). Data assimilation of satellite-based soil moisture into a distributed hydrological model for streamflow predictions. *Hydrology*, *8*(1), 52. https://doi.org/10.3390/hydrology8010052

Jasper, K., Gurtz, J., & Lang, H. (2002). Advanced flood forecasting in Alpine watersheds by coupling meteorological observations and forecasts with a distributed hydrological model. *Journal of Hydrology*, *267*(1-2), 40-52. https://doi.org/10.1016/S0022-1694(02)00138-5

Jeong, D. I., & Kim, Y. O. (2005). Rainfall-runoff models using artificial neural networks for ensemble streamflow prediction. *Hydrological Processes: An International Journal*, *19*(19), 3819-3835. https://doi.org/10.1002/hyp.5983

Jiménez, K. Q., Collischonn, W., & Paiva, R. C. D. d. (2019). Data assimilation using the ensemble Kalman filter in a distributed hydrological model on the Tocantins River, Brasil. *RBRH*, *24*. https://doi.org/10.1590/2318-0331.241920180031

Kalra, A., Miller, W. P., Lamb, K. W., Ahmad, S., & Piechota, T. (2013). Using large-scale climatic patterns for improving long lead time streamflow forecasts for Gunnison and San Juan River Basins. *Hydrological Processes*, *27*(11), 1543-1559. https://doi.org/10.1002/hyp.9236

Kang, T. H., Kim, Y. O., & Hong, I. P. (2010). Comparison of pre-and post-processors for ensemble streamflow prediction. *Atmospheric science letters*, *11*(2), 153-159. https://doi.org/10.1002/asl.276

Kao, I.-F., Liou, J.-Y., Lee, M.-H., & Chang, F.-J. (2021). Fusing stacked autoencoder and long short-term memory for regional multistep-ahead flood inundation forecasts. *Journal of Hydrology*, *598*, 126371. https://doi.org/10.1016/j.jhydrol.2021.126371

Kao, I.-F., Zhou, Y., Chang, L.-C., & Chang, F.-J. (2020). Exploring a Long Short-Term Memory based Encoder-Decoder framework for multi-step-ahead flood forecasting. *Journal of Hydrology*, *583*, 124631. https://doi.org/10.1016/j.jhydrol.2020.124631

Khaki, M., Hoteit, I., Kuhn, M., Awange, J., Forootan, E., Van Dijk, A. I. J. M., Schumacher, M., & Pattiaratchi, C. (2017). Assessing sequential data assimilation techniques for integrating GRACE data into a hydrological model. *Advances in Water Resources*, *107*, 301-316. https://doi.org/10.1016/j.advwatres.2017.07.001

Khoshkalam, Y., Rousseau, A. N., Rahmani, F., Shen, C., & Abbasnezhadi, K. (2023). Applying Transfer Learning Techniques to Enhance the Accuracy of Streamflow Prediction Produced by Long Short-term Memory Networks with Data Integration. *Journal of Hydrology*, 129682. https://doi.org/10.1016/j.jhydrol.2023.129682

Kim, K.-J., Kim, Y.-O., & Kang, T.-H. (2017). Application of time-lagged ensemble approach with auto-regressive processors to reduce uncertainties in peak discharge and timing. *Journal of Hydrology: Regional Studies*, *9*, 140-148. https://doi.org/10.1016/j.ejrh.2016.12.081

Kim, K. B., Kwon, H.-H., & Han, D. (2018). Exploration of warm-up period in conceptual hydrological modelling. *Journal of Hydrology*, *556*, 194-210. https://doi.org/10.1016/j.jhydrol.2017.11.015

Komma, J., Reszler, C., Blöschl, G., & Haiden, T. (2007). Ensemble prediction of floods–catchment non-linearity and forecast probabilities. *Natural Hazards and Earth System Sciences*, *7*(4), 431-444. https://doi.org/10.5194/nhess-7-431-2007

Konapala, G., Valiya Veettil, A., & Mishra, A. K. (2018). Teleconnection between low flows and large-scale climate indices in Texas River basins. *Stochastic Environmental*

*Research and Risk Assessment*, *32*(8), 2337-2350. https://doi.org/10.1007/s00477-017-1460-6

Koutsoyiannis, D., Yao, H., & Georgakakos, A. (2008). Medium-range flow prediction for the Nile: a comparison of stochastic and deterministic methods/Prévision du débit du Nil à moyen terme: une comparaison de méthodes stochastiques et déterministes. *Hydrological Sciences Journal*, *53*(1), 142-164. https://doi.org/10.1623/hysj.53.1.142

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. *Hydrology and Earth System Sciences*, *22*(11), 6005-6022. https://doi.org/10.5194/hess-22-6005-2018

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. *Hydrology and Earth System Sciences*, *23*(12), 5089-5110. https://doi.org/10.5194/hess-23-5089-2019

Kroll, C., Luz, J., Allen, B., & Vogel, R. M. (2004). Developing a watershed characteristics database to improve low streamflow prediction. *Journal of Hydrologic Engineering*, *9*(2), 116-125. https://doi.org/10.1061/(ASCE)1084-0699(2004)9:2(116)

Krzysztofowicz, R. (1999). Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resources Research*, *35*(9), 2739-2750. https://doi.org/10.1029/1999WR900099

Krzysztofowicz, R. (2001). The case for probabilistic forecasting in hydrology. *Journal of Hydrology*, *249*(1-4), 2-9. https://doi.org/10.1016/S0022-1694(01)00420-6

Lamb, K. W., Piechota, T. C., Aziz, O. A., & Tootle, G. A. (2011). Basis for extending long-term streamflow forecasts in the Colorado River basin. *Journal of Hydrologic Engineering*, *16*(12), 1000-1008. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000153

Le, Ho, Lee, & Jung. (2019). Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting. *Water*, *11*(7), 1387. https://doi.org/10.3390/w11071387

Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., & Dadson, S. J. (2021). Benchmarking data-driven rainfall–runoff models in Great Britain: a comparison of long short-term memory (LSTM)-based models with four lumped conceptual models. *Hydrology and Earth System Sciences*, *25*(10), 5517-5534. https://doi.org/10.5194/hess-25-5517-2021

Li, H., Luo, L., Wood, E. F., & Schaake, J. (2009). The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting. *Journal of Geophysical Research: Atmospheres*, *114*(D4). https://doi.org/10.1029/2008JD010969

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., & Di, Z. (2017). A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *Wiley Interdisciplinary Reviews: Water*, *4*(6), e1246. https://doi.org/10.1002/wat2.1246

Li, W., Kiaghadi, A., & Dawson, C. (2021a). Exploring the best sequence LSTM modeling architecture for flood prediction. *Neural Computing and Applications*, *33*, 5571-5580. https://doi.org/10.1007/s00521-020-05334-3

Li, W., Kiaghadi, A., & Dawson, C. (2021b). High temporal resolution rainfall–runoff modeling using long-short-term-memory (LSTM) networks. *Neural Computing and Applications*, *33*, 1261-1278. https://doi.org/10.1007/s00521-020-05010-6

Li, W., & Sankarasubramanian, A. (2012). Reducing hydrologic model uncertainty in monthly streamflow predictions using multimodel combination. *Water Resources Research*, *48*(12). https://doi.org/10.1029/2011WR011380

Liu, B., Tang, Q., Zhao, G., Gao, L., Shen, C., & Pan, B. (2022). Physics-guided long short-term memory network for streamflow and flood simulations in the Lancang–Mekong river basin. *Water*, *14*(9), 1429. https://doi.org/10.3390/w14091429

Liu, Y., & Gupta, H. V. (2007). Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. *Water Resources Research*, *43*(7). https://doi.org/10.1029/2006WR005756

Liu, Y., Weerts, A., Clark, M., Hendricks Franssen, H.-J., Kumar, S., Moradkhani, H., Seo, D.-J., Schwanenberg, D., Smith, P., & Van Dijk, A. (2012). Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities. *Hydrology and Earth System Sciences*, *16*(10), 3863-3887. https://doi.org/10.5194/hess-16-3863-2012

Loiselle, G., Martel, J. L., Poulin, A., Lachance-Cloutier, S., Turcotte, R., Fournier, J., Mai, J., & Arsenault, R. (2021). A semi-empirical wind set-up forecasting model for Lake Champlain. *Hydrological Processes*, *35*(6), e14240. https://doi.org/10.1002/hyp.14240

Lucatero, D., Madsen, H., Refsgaard, J. C., Kidmose, J., & Jensen, K. H. (2018). Seasonal streamflow forecasts in the Ahlergaarde catchment, Denmark: the effect of preprocessing and post-processing on skill and statistical consistency. *Hydrology and Earth System Sciences*, *22*(7), 3601-3617. https://doi.org/10.5194/hess-22-3601-2018

Machado, F., Mine, M., Kaviski, E., & Fill, H. (2011). Monthly rainfall–runoff modelling using artificial neural networks. *Hydrological Sciences Journal–Journal des Sciences Hydrologiques*, *56*(3), 349-361. https://doi.org/10.1080/02626667.2011.559949

MacLachlan, C., Arribas, A., Peterson, K. A., Maidens, A., Fereday, D., Scaife, A., Gordon, M., Vellinga, M., Williams, A., & Comer, R. (2015). Global Seasonal forecast system version 5 (GloSea5): A high-resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society*, *141*(689), 1072-1084. https://doi.org/10.1002/qj.2396

Madadgar, S., & Moradkhani, H. (2013). A Bayesian framework for probabilistic seasonal drought forecasting. *Journal of Hydrometeorology*, *14*(6), 1685-1705. https://doi.org/10.1175/JHM-D-13-010.1

Madsen, H. (2003). Forecasting system using Kalman filtering. *Water Resources Systems: Hydrological Risk, Management and Development*(278), 75.

Mai, J., Arsenault, R., Tolson, B. A., Latraverse, M., & Demeester, K. (2020). Application of parameter screening to derive optimal initial state adjustments for streamflow forecasting. *Water Resources Research*, *56*(9), e2020WR027960. https://doi.org/10.1029/2020WR027960

Maier, H. R., Jain, A., Dandy, G. C., & 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 & software*, *25*(8), 891-909. https://doi.org/10.1016/j.envsoft.2010.02.003

Maity, R., & Nagesh Kumar, D. (2008). Basin-scale stream-flow forecasting using the information of large-scale atmospheric circulation phenomena. *Hydrological Processes: An International Journal*, *22*(5), 643-650. https://doi.org/10.1002/hyp.6630

Mariotti, A., Barnes, E., Chang, E., Lang, A., Dirmeyer, P., Pegion, K., Barrie, D., & Baggett, C. (2019). Bridging the weather-to-climate prediction gap. https://doi.org/10.1029/2019EO115819

Martel, J.-L., Demeester, Kenjy, Brissette, François, Poulin, Annie et Arsenault, Richard. (2017). HMETS—A simple and efficient hydrology model for teaching hydrological modelling, flow forecasting and climate change impacts. *International Journal of Engineering Education*, *33*, 1307-1316. https://dialnet.unirioja.es/servlet/articulo?codigo=6897050

Mather, A. L., & Johnson, R. L. (2016). Forecasting Turbidity during Streamflow Events for Two Mid-Atlantic U.S. Streams. *Water Resources Management*, *30*(13), 4899-4912. https://doi.org/10.1007/s11269-016-1460-1

Matheson, J. E., & Winkler, R. L. (1976). Scoring rules for continuous probability distributions. *Management science*, *22*(10), 1087-1096.

Maurer, E. P., & Lettenmaier, D. P. (2004). Potential effects of long-lead hydrologic predictability on Missouri River main-stem reservoirs. *Journal of Climate*, *17*(1), 174-186. https://doi.org/10.1175/1520-0442(2004)017%3C0174:PEOLHP%3E2.0.CO;2

Mazzoleni, M., Noh, S. J., Lee, H., Liu, Y., Seo, D.-J., Amaranto, A., Alfonso, L., & Solomatine, D. P. (2018). Real-time assimilation of streamflow observations into a hydrological routing model: effects of model structures and updating methods. *Hydrological Sciences Journal*, *63*(3), 386-407. https://doi.org/10.1080/02626667.2018.1430898

McCabe, G. J., Palecki, M. A., & Betancourt, J. L. (2004). Pacific and Atlantic Ocean influences on multidecadal drought frequency in the United States. *Proceedings of the National Academy of Sciences*, *101*(12), 4136-4141. https://doi.org/10.1073/pnas.0306738101

McInerney, D., Thyer, M., Kavetski, D., Laugesen, R., Tuteja, N., & Kuczera, G. (2020). Multi-temporal hydrological residual error modeling for seamless subseasonal streamflow forecasting. *Water Resources Research*, *56*(11), e2019WR026979. https://doi.org/10.1029/2019WR026979

McInerney, D., Thyer, M., Kavetski, D., Laugesen, R., Woldemeskel, F., Tuteja, N., & Kuczera, G. (2022). Seamless streamflow forecasting at daily to monthly scales: MuTHRE lets you have your cake and eat it too. *Hydrology and Earth System Sciences*, *26*(21), 5669-5683. https://doi.org/10.5194/hess-26-5669-2022

McMillan, H., Hreinsson, E., Clark, M., Singh, S., Zammit, C., & Uddstrom, M. (2013). Operational hydrological data assimilation with the recursive ensemble Kalman filter. *Hydrology and Earth System Sciences*, *17*(1), 21-38. https://doi.org/10.5194/hess-17-21-2013

Mehedi, M. A. A., Khosravi, M., Yazdan, M. M. S., & Shabanian, H. (2022). Exploring Temporal Dynamics of River Discharge using Univariate Long Short-Term Memory (LSTM) Recurrent Neural Network at East Branch of Delaware River. *Hydrology*, *9*(11), 202. https://doi.org/10.3390/hydrology9110202

Mendoza, P. A., Rajagopalan, B., Clark, M. P., Cortés, G., & McPhee, J. (2014). A robust multimodel framework for ensemble seasonal hydroclimatic forecasts. *Water Resources Research*, *50*(7), 6030-6052. https://doi.org/10.1002/2014WR015426

Mendoza, P. A., Wood, A. W., Clark, E., Rothwell, E., Clark, M. P., Nijssen, B., Brekke, L. D., & Arnold, J. R. (2017). An intercomparison of approaches for improving operational seasonal streamflow forecasts. *Hydrology and Earth System Sciences*, *21*(7), 3915-3935. https://doi.org/10.5194/hess-2017-60

Ming, X., Liang, Q., Xia, X., Li, D., & Fowler, H. J. (2020). Real-time flood forecasting based on a high-performance 2-D hydrodynamic model and numerical weather predictions. *Water Resources Research*, *56*(7), e2019WR025583. https://doi.org/10.1029/2019WR025583

Monhart, S., Zappa, M., Spirig, C., Schär, C., & Bogner, K. (2019). Subseasonal hydrometeorological ensemble predictions in small-and medium-sized mountainous catchments: benefits of the NWP approach. *Hydrology and Earth System Sciences*, *23*(1), 493-513. https://doi.org/10.5194/hess-23-493-2019

Moradkhani, H., Nearing, G., Abbaszadeh, P., & Pathiraja, S. (2018). Fundamentals of data assimilation and theoretical advances. *Handbook of hydrometeorological ensemble forecasting*, 1-26. https://doi.org/10.1007/978-3-642-39925-1

Moradkhani, H., & Sorooshian, S. (2009). General review of rainfall-runoff modeling: model calibration, data assimilation, and uncertainty analysis. *Hydrological modelling and the water cycle*, 1-24. https://doi.org/10.1007/978-3-540-77843-1_1

Moradkhani, H., Sorooshian, S., Gupta, H. V., & Houser, P. R. (2005). Dual state–parameter estimation of hydrological models using ensemble Kalman filter. *Advances in Water Resources*, *28*(2), 135-147. https://doi.org/10.1016/j.advwatres.2004.09.002

Morin, G., & Paquet, P. (2007). Modèle hydrologique CEQUEAU.

Mudelsee, M., Börngen, M., Tetzlaff, G., & Grünewald, U. (2003). No upward trends in the occurrence of extreme floods in central Europe. *nature*, *425*(6954), 166-169. https://doi.org/10.1038/nature01928

Muluye, G. Y. (2011). Improving long-range hydrological forecasts with extended Kalman filters. *Hydrological Sciences Journal*, *56*(7), 1118-1128. https://doi.org/10.1080/02626667.2011.608068

Najafi, M. R., Moradkhani, H., & Piechota, T. C. (2012). Ensemble streamflow prediction: climate signal weighting methods vs. climate forecast system reanalysis. *Journal of Hydrology*, *442*, 105-116. https://doi.org/10.1016/j.jhydrol.2012.04.003

Nearing, G. S., Klotz, D., Frame, J. M., Gauch, M., Gilon, O., Kratzert, F., Sampson, A. K., Shalev, G., & Nevo, S. (2022). Data assimilation and autoregression for using near-real-time streamflow observations in long short-term memory networks. *Hydrology*

*and Earth System Sciences*, *26*(21), 5493-5513. https://doi.org/10.5194/hess-26-5493-2022

Nester, T., Komma, J., Viglione, A., & Blöschl, G. (2012). Flood forecast errors and ensemble spread—A case study. *Water Resources Research*, *48*(10). https://doi.org/10.1029/2011WR011649

Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, D., Voloshin, D., Kratzert, F., Elidan, G., & Dror, G. (2022). Flood forecasting with machine learning models in an operational framework. *Hydrology and Earth System Sciences*, *26*(15), 4013-4032. https://doi.org/10.5194/hess-26-4013-2022

Noh, S. J., Weerts, A. H., Rakovec, O., Lee, H., & Seo, D.-J. (2018). Assimilation of streamflow observations. In *Handbook of hydrometeorological ensemble forecasting* (pp. 1-36). Springer Verlag. https://doi.org/10.3390/en16010237

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the earth sciences. *Science*, *263*(5147), 641-646. https://doi.org/10.1126/science.263.5147.641

Oudin, L., Hervieu, F., Michel, C., Perrin, C., Andréassian, V., Anctil, F., & Loumagne, C. (2005). Which potential evapotranspiration input for a lumped rainfall–runoff model?: Part 2—Towards a simple and efficient potential evapotranspiration model for rainfall–runoff modelling. *Journal of Hydrology*, *303*(1-4), 290-306. https://doi.org/10.1016/j.jhydrol.2004.08.026

Pagano, T., Hapuarachchi, P., & Wang, Q. (2010). Continuous rainfall-runoff model comparison and short-term daily streamflow forecast skill evaluation. https://doi.org/10.4225/08/58542c672dd2c

Pappenberger, F., Beven, K. J., Hunter, N., Bates, P., Gouweleeuw, B. T., Thielen, J., & De Roo, A. (2005). Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European Flood Forecasting System (EFFS). *Hydrology and Earth System Sciences*, *9*(4), 381-393. https://doi.org/10.5194/hess-9-381-2005

Pappenberger, F., Jendritzky, G., Staiger, H., Dutra, E., Di Giuseppe, F., Richardson, D., & Cloke, H. (2015). Global forecasting of thermal health hazards: the skill of probabilistic predictions of the Universal Thermal Climate Index (UTCI). *International journal of biometeorology*, *59*(3), 311-323. https://doi.org/10.1007/s00484-014-0843-3

Patil, A., & Ramsankaran, R. (2017). Improving streamflow simulations and forecasting performance of SWAT model by assimilating remotely sensed soil moisture

observations. *Journal of Hydrology*, *555*, 683-696. https://doi.org/10.1016/j.jhydrol.2017.10.058

Piazzi, G., Thirel, G., Perrin, C., & Delaigue, O. (2021). Sequential data assimilation for streamflow forecasting: assessing the sensitivity to uncertainties and updated variables of a conceptual hydrological model at basin scale. *Water Resources Research*, *57*(4). https://doi.org/10.1029/2020WR028390

Ponnoprat, D. (2021). Short-term daily precipitation forecasting with seasonally-integrated autoencoder. *Applied Soft Computing*, *102*, 107083. https://doi.org/10.1016/j.asoc.2021.107083

Prakash, V., & Mishra, V. (2022). Soil moisture and streamflow data assimilation for streamflow prediction in the Narmada River Basin. *Journal of Hydrometeorology*. https://doi.org/10.1175/JHM-D-21-0139.1

Provotar, O. I., Linder, Y. M., & Veres, M. M. (2019). Unsupervised anomaly detection in time series using lstm-based autoencoders. 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT), https://doi.org/10.1109/ATIT49449.2019.9030505

Prudhomme, C., Hannaford, J., Harrigan, S., Boorman, D., Knight, J., Bell, V., Jackson, C., Svensson, C., Parry, S., & Bachiller-Jareno, N. (2017). Hydrological Outlook UK: an operational streamflow and groundwater level forecasting system at monthly to seasonal time scales. *Hydrological Sciences Journal*, *62*(16), 2753-2768. https://doi.org/10.1080/02626667.2017.1395032

Quedi, E. S., & Fan, F. M. (2020). Sub seasonal streamflow forecast assessment at large-scale basins. *Journal of Hydrology*, *584*, 124635. https://doi.org/10.1016/j.jhydrol.2020.124635

Rajagopalan, B., Salas, J. D., & Lall, U. (2010). Stochastic methods for modeling precipitation and streamflow. In *Advances in data-based approaches for hydrologic modeling and forecasting* (pp. 17-52). World Scientific. https://doi.org/10.1142/9789814307987_0002

Rakovec, O., Weerts, A., Hazenberg, P., Torfs, P., & Uijlenhoet, R. (2012). State updating of a distributed hydrological model with Ensemble Kalman Filtering: effects of updating frequency and observation network density on forecast accuracy. *Hydrology and Earth System Sciences*, *16*(9), 3435-3449. https://doi.org/10.5194/hess-16-3435-2012

Rakovec, O., Weerts, A., Sumihar, J., & Uijlenhoet, R. (2015). Operational aspects of asynchronous filtering for flood forecasting. *Hydrology and Earth System Sciences*, *19*(6), 2911-2924. https://doi.org/10.5194/hess-19-2911-2015

Ramos, M.-H., Castelletti, A., Pulido-Velazquez, M., & Gustafsson, D. (2016). Weather and climate services for hydropower management. Hydropower and Environmental Sustainability,

Ramos, M. H., Van Andel, S. J., & Pappenberger, F. (2013). Do probabilistic forecasts lead to better decisions? *Hydrology and Earth System Sciences*, *17*(6), 2219-2232. https://doi.org/10.5194/hess-17-2219-2013

Reichle, R. H. (2008). Data assimilation methods in the Earth sciences. *Advances in Water Resources*, *31*(11), 1411-1418. https://doi.org/10.1016/j.advwatres.2008.01.001

Robertson, D. E., & Wang, Q. (2012). A Bayesian approach to predictor selection for seasonal streamflow forecasting. *Journal of Hydrometeorology*, *13*(1), 155-171. https://doi.org/10.1175/JHM-D-10-05009.1

Rodda, J. C. (2011). Guide to Hydrological Practices: vol. I: Hydrology—From Measurement to Hydrological Information, and vol. II: Management of Water Resources and Application to Hydrological Practices, 2008 and 2009, WMO 168, World Meteorological Organziation, Geneva, Switzerland; Price CHF70. 00; ISBN 978-92-63-10168-6. In: Taylor & Francis.

Roulin, E. (2007). Skill and relative economic value of medium-range hydrological ensemble predictions. *Hydrology and Earth System Sciences*, *11*(2), 725-737. https://doi.org/10.5194/hess-11-725-2007

Roundy, J. K., Duan, Q., & Schaake, J. C. (2019). Hydrological predictability, scales, and uncertainty issues. In *Handbook of hydrometeorological ensemble forecasting* (pp. 3-31). Springer. https://doi.org/10.1007/978-3-642-39925-1_8

Roy, T., Serrat-Capdevila, A., Gupta, H., & Valdes, J. (2017). A platform for probabilistic Multimodel and Multiproduct Streamflow Forecasting. *Water Resources Research*, *53*(1), 376-399. https://doi.org/10.1002/2016WR019752

Saavedra Valeriano, O. C., Koike, T., Yang, K., Graf, T., Li, X., Wang, L., & Han, X. (2010). Decision support for dam release during floods using a distributed biosphere hydrological model driven by quantitative precipitation forecasts. *Water Resources Research*, *46*(10). https://doi.org/10.1029/2010WR009502

Sagarika, S., Kalra, A., & Ahmad, S. (2014). Evaluating the effect of persistence on long-term trends and analyzing step changes in streamflows of the continental United States. *Journal of Hydrology*, *517*, 36-53. https://doi.org/10.1016/j.jhydrol.2014.05.002

Sahoo, B. B., Jha, R., Singh, A., & Kumar, D. (2019). Long short-term memory (LSTM) recurrent neural network for low-flow hydrological time series forecasting. *Acta Geophysica*, *67*(5), 1471-1481. https://doi.org/10.1007/s11600-019-00330-1

Samuel, J., Rousseau, A. N., Abbasnezhadi, K., & Savary, S. (2019). Development and evaluation of a hydrologic data-assimilation scheme for short-range flow and inflow forecasts in a data-sparse high-latitude region using a distributed model and ensemble Kalman filtering. *Advances in Water Resources*, *130*, 198-220. https://doi.org/10.1016/j.advwatres.2019.06.004

Schaake, J., Demargne, J., Hartman, R., Mullusky, M., Welles, E., Wu, L., Herr, H., Fan, X., & Seo, D. (2007). Precipitation and temperature ensemble forecasts from single-value forecasts. *Hydrology and Earth System Sciences Discussions*, *4*(2), 655-717. https://doi.org/10.5194/hessd-4-655-2007

Schaake, J., Franz, K., Bradley, A., & Buizza, R. (2006). The hydrologic ensemble prediction experiment (HEPEX). *Hydrology and Earth System Sciences Discussions*, *3*(5), 3321-3332. https://doi.org/10.5194/hessd-3-3321-2006

Schmidhuber, J., & Hochreiter, S. (1997). Long short-term memory. *Neural Comput*, *9*(8), 1735-1780. https://doi.org/https://dx.doi.org/10.1162/neco.1997.9.8.1735

Schwanenberg, D., Fan, F. M., Naumann, S., Kuwajima, J. I., Montero, R. A., & Assis dos Reis, A. (2015). Short-term reservoir optimization for flood mitigation under meteorological and hydrological forecast uncertainty. *Water Resources Management*, *29*(5), 1635-1651. https://doi.org/10.1007/s11269-014-0899-1

Seiller, G., & Anctil, F. (2014). Climate change impacts on the hydrologic regime of a Canadian river: comparing uncertainties arising from climate natural variability and lumped hydrological model structures. *Hydrology and Earth System Sciences*, *18*(6), 2033-2047. https://doi.org/10.5194/hess-18-2033-2014

Şensoy, A., & Uysal, G. (2012). The value of snow depletion forecasting methods towards operational snowmelt runoff estimation using MODIS and Numerical Weather Prediction Data. *Water Resources Management*, *26*(12), 3415-3440. https://doi.org/10.1007/s11269-012-0079-0

Seo, D.-J., Herr, H., & Schaake, J. (2006). A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. *Hydrology and Earth System Sciences Discussions*, *3*(4), 1987-2035. https://doi.org/10.5194/hessd-3-1987-2006

Shams, M. S., Anwar, A. F., Lamb, K. W., & Bari, M. (2018). Relating ocean-atmospheric climate indices with Australian river streamflow. *Journal of Hydrology*, *556*, 294-309. https://doi.org/10.1016/j.jhydrol.2017.11.017

Shen, C., & Lawson, K. (2021). Applications of deep learning in hydrology. *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences*, 283-297. https://doi.org/10.1002/9781119646181.ch19

Shi, H., Li, T., Liu, R., Chen, J., Li, J., Zhang, A., & Wang, G. (2015). A service-oriented architecture for ensemble flood forecast from numerical weather prediction. *Journal of Hydrology*, *527*, 933-942. https://doi.org/10.1016/j.jhydrol.2015.05.056

Shrestha, D., Robertson, D., Wang, Q., Pagano, T., & Hapuarachchi, H. (2013). Evaluation of numerical weather prediction model precipitation forecasts for short-term streamflow forecasting purpose. *Hydrology and Earth System Sciences*, *17*(5), 1913-1931. https://doi.org/10.5194/hess-17-1913-2013

Smiatek, G., Kunstmann, H., & Werhahn, J. (2012). Implementation and performance analysis of a high resolution coupled numerical weather and river runoff prediction model system for an Alpine catchment. *Environmental modelling & software*, *38*, 231-243. https://doi.org/10.1016/j.envsoft.2012.06.001

Sohrabi, S., & Brissette, F. P. (2021). Evaluation of a stochastic weather generator for long-term ensemble streamflow forecasts. *Hydrological Sciences Journal*, *66*(3), 474-487. https://doi.org/10.1080/02626667.2021.1873343

Sohrabi, S., Brissette, F. P., & Arsenault, R. (2021). Coupling large-scale climate indices with a stochastic weather generator to improve long-term streamflow forecasts in a Canadian watershed. *Journal of Hydrology*, *594*, 125925. https://doi.org/10.1016/j.jhydrol.2020.125925

Solomatine, D., See, L. M., & Abrahart, R. (2009). Data-driven modelling: concepts, approaches and experiences. *Practical hydroinformatics*, 17-30. https://doi.org/10.1007/978-3-540-79881-1_2

Sorooshian, S., Duan, Q., & Gupta, V. K. (1993). Calibration of rainfall-runoff models: Application of global optimization to the Sacramento Soil Moisture Accounting Model. *Water Resources Research*, *29*(4), 1185-1194. https://doi.org/10.1029/92WR02617

Soukup, T. L., Aziz, O. A., Tootle, G. A., Piechota, T. C., & Wulff, S. S. (2009). Long lead-time streamflow forecasting of the North Platte River incorporating oceanic–

atmospheric climate variability. *Journal of Hydrology*, *368*(1-4), 131-142. https://doi.org/10.1016/j.jhydrol.2008.11.047

Stern, H., & Davidson, N. E. (2015). Trends in the skill of weather prediction at lead times of 1–14 days. *Quarterly Journal of the Royal Meteorological Society*, *141*(692), 2726-2736. https://doi.org/10.1002/qj.2559

Sun, L., Seidou, O., Nistor, I., & Liu, K. (2016). Review of the Kalman-type hydrological data assimilation. *Hydrological Sciences Journal*, *61*(13), 2348-2366. https://doi.org/10.1080/02626667.2015.1127376

Sun, R., Yuan, H., & Yang, Y. (2018). Using multiple satellite-gauge merged precipitation products ensemble for hydrologic uncertainty analysis over the Huaihe River basin. *Journal of Hydrology*, *566*, 406-420. https://doi.org/10.1016/j.jhydrol.2018.09.024

Svensson, C. (2016). Seasonal river flow forecasts for the United Kingdom using persistence and historical analogues. *Hydrological Sciences Journal*, *61*(1), 19-35. https://doi.org/10.1080/02626667.2014.992788

Thiboult, A., & Anctil, F. (2015). On the difficulty to optimally implement the Ensemble Kalman filter: An experiment based on many hydrological models and catchments. *Journal of Hydrology*, *529*, 1147-1160. https://doi.org/10.1016/j.jhydrol.2015.09.036

Thiboult, A., Anctil, F., & Boucher, M.-A. (2016). Accounting for three sources of uncertainty in ensemble hydrological forecasting. *Hydrology and Earth System Sciences*, *20*(5), 1809-1825. https://doi.org/10.5194/hessd-12-7179-2015

Thiboult, A., Seiller, G., Poncelet, C., & Anctil, F. (2020). The HOOPLA toolbox: a HydrOlOgical Prediction LAboratory to explore ensemble rainfall-runoff modeling. *Hydrology and Earth System Sciences Discussions*, 1-18. https://doi.org/10.5194/hess-2020-6

Thielen-del Pozo, J., & Bruen, M. (2019). Overview of forecast communication and use of ensemble hydrometeorological forecasts. *Handbook of hydrometeorological ensemble forecasting*, 1037-1045. https://doi.org/10.1007/978-3-642-40457-3_40-1

Trenberth, K. (1997). Bulletin of the American Meteorological Society. *The definition of el nino*, *78*(12), 2771-2778. https://doi.org/10.1175/1520-0477(1997)078%3C2771:TDOENO%3E2.0.CO;2

Troin, M., Arsenault, R., & Brissette, F. (2015). Performance and uncertainty evaluation of snow models on snowmelt flow simulations over a Nordic catchment (Mistassibi, Canada). *Hydrology*, *2*(4), 289-317. https://doi.org/10.3390/hydrology2040289

Troin, M., Arsenault, R., Wood, A. W., Brissette, F., & Martel, J. L. (2021). Generating Ensemble Streamflow Forecasts: A Review of Methods and Approaches Over the Past 40 Years. *Water Resources Research*, *57*(7). https://doi.org/10.1029/2020wr028392

Twedt, T. M., Schaake Jr, J. C., & Peck, E. L. (1977). National Weather Service extended streamflow prediction [USA]. Proceedings Western Snow Conference,

Valdez, E. S., Anctil, F., & Ramos, M.-H. (2022). Choosing between post-processing precipitation forecasts or chaining several uncertainty quantification tools in hydrological forecasting systems. *Hydrology and Earth System Sciences*, *26*(1), 197-220. https://doi.org/10.5194/hess-26-197-2022

Valéry, A. (2010). *Modélisation précipitations débit sous influence nivale: Elaboration d'un module neige et évaluation sur 380 bassins versants* Doctorat Hydrobiologie, Institut des Sciences et Industries du Vivant et de …]. https://webgr.irstea.fr/wp-content/uploads/2012/07/2010-VALERY-THESE.pdf

Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., & Atencia, A. (2020). Statistical postprocessing for weather forecasts–review, challenges and avenues in a big data world. *Bulletin of the American Meteorological Society*, 1-44. https://doi.org/10.1175/BAMS-D-19-0308.1

Vergara, H., Hong, Y., & Gourley, J. (2014). Improving Flood Forecasting Skill with the Ensemble Kalman Filter. *Revista de Tecnología*, *13*(1), 9-27. https://doi.org/10.18270/rt.v13i1.1294

Verkade, J., Brown, J., Reggiani, P., & Weerts, A. (2013). Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *Journal of Hydrology*, *501*, 73-91. https://doi.org/10.1016/j.jhydrol.2013.07.039

Vitart, F., & Robertson, A. W. (2018). The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *npj Climate and Atmospheric Science*, *1*(1), 1-7. https://doi.org/10.1038/s41612-018-0013-0

Wang, H., Chen, Y., & Chen, Z. (2013). Spatial distribution and temporal trends of mean precipitation and extremes in the arid region, northwest of China, during 1960–2010. *Hydrological Processes*, *27*(12), 1807-1818. https://doi.org/10.1002/hyp.9339

Wang, Y., Yao, H., & Zhao, S. (2016). Auto-encoder based dimensionality reduction. *Neurocomputing*, *184*, 232-242. https://doi.org/10.1016/j.neucom.2015.08.104

Weerts, A. H., & El Serafy, G. Y. (2006). Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models. *Water Resources Research*, *42*(9). https://doi.org/10.1029/2005WR004093

Werner, K., Brandon, D., Clark, M., & Gangopadhyay, S. (2004). Climate index weighting schemes for NWS ESP-based seasonal volume forecasts. *Journal of Hydrometeorology*, *5*(6), 1076-1090. https://doi.org/10.1175/JHM-381.1

Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and computing*, *4*(2), 65-85. https://doi.org/10.1007/BF00175354

WMO. (2009). Guide to hydrological practices, volume II management of water resources and application of hydrological practices. In: WMO Geneva, Switzerland.

Woldemeskel, F., McInerney, D., Lerat, J., Thyer, M., Kavetski, D., Shin, D., Tuteja, N., & Kuczera, G. (2018). Evaluating post-processing approaches for monthly and seasonal streamflow forecasts. *Hydrology and Earth System Sciences*, *22*(12), 6257-6278. https://doi.org/10.5194/hess-22-6257-2018

Wood, A. W., Hopson, T., Newman, A., Brekke, L., Arnold, J., & Clark, M. (2016). Quantifying streamflow forecast skill elasticity to initial condition and climate prediction skill. *Journal of Hydrometeorology*, *17*(2), 651-668. https://doi.org/10.1175/JHM-D-14-0213.1

Wu, C., Chau, K. W., & Li, Y. S. (2009). Predicting monthly streamflow using data-driven models coupled with data-preprocessing techniques. *Water Resources Research*, *45*(8). https://doi.org/10.1029/2007WR006737

Xie, X., & Zhang, D. (2010). Data assimilation for distributed hydrological catchment modeling via ensemble Kalman filter. *Advances in Water Resources*, *33*(6), 678-690. https://doi.org/10.1016/j.advwatres.2010.03.012

Xu, W., Jiang, Y., Zhang, X., Li, Y., Zhang, R., & Fu, G. (2020). Using long short-term memory networks for river flow prediction. *Hydrology Research*, *51*(6), 1358-1376. https://doi.org/10.2166/nh.2020.026

Yang, C., Yuan, H., & Su, X. (2020). Bias correction of ensemble precipitation forecasts in the improvement of summer streamflow prediction skill. *Journal of Hydrology*, *588*, 124955. https://doi.org/10.1016/j.jhydrol.2020.124955

Yao, H., & Georgakakos, A. (2001). Assessment of Folsom Lake response to historical and potential future climate scenarios: 2. Reservoir management. *Journal of Hydrology*, *249*(1-4), 176-196. https://doi.org/10.1016/S0022-1694(01)00418-8

Yuan, X., Wood, E. F., & Liang, M. (2014). Integrating weather and climate prediction: Toward seamless hydrologic forecasting. *Geophysical Research Letters*, *41*(16), 5891-5896. https://doi.org/10.1002/2014GL061076

Yuan, X., Wood, E. F., & Ma, Z. (2015). A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development. *Wiley Interdisciplinary Reviews: Water*, *2*(5), 523-536. https://doi.org/10.1002/wat2.1088

Yucel, I., Onen, A., Yilmaz, K., & Gochis, D. (2015). Calibration and evaluation of a flood forecasting system: Utility of numerical weather prediction model, data assimilation and satellite-based rainfall. *Journal of Hydrology*, *523*, 49-66. https://doi.org/10.1016/j.jhydrol.2015.01.042

Zalachori, I., Ramos, M.-H., Garçon, R., Mathevet, T., & Gailhard, J. (2012). Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies. *Advances in Science and Research*, *8*(1), 135-141. https://doi.org/10.5194/asr-8-135-2012

Zamo, M., & Naveau, P. (2018). Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts. *Mathematical Geosciences*, *50*(2), 209-234. https://doi.org/10.1007/s11004-017-9709-7

Zappa, M., Beven, K. J., Bruen, M., Cofino, A. S., Kok, K., Martin, E., Nurmi, P., Orfila, B., Roulin, E., & Schröter, K. (2010). Propagation of uncertainty from observing systems and NWP into hydrological models: COST-731 Working Group 2. *Atmospheric science letters*, *11*(2), 83-91. https://doi.org/10.1002/asl.248

Zhang, J., Chen, J., Li, X., Chen, H., Xie, P., & Li, W. (2020). Combining postprocessed ensemble weather forecasts and multiple hydrological models for ensemble streamflow predictions. *Journal of Hydrologic Engineering*, *25*(1), 04019060. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001871

Zhang, J., Zhu, Y., Zhang, X., Ye, M., & Yang, J. (2018). Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. *Journal of Hydrology*, *561*, 918-929. https://doi.org/10.1016/j.jhydrol.2018.04.065

Zhang, X., Peng, Y., Zhang, C., & Wang, B. (2015). Are hybrid models integrated with data preprocessing techniques suitable for monthly streamflow forecasting? Some experiment evidences. *Journal of Hydrology*, *530*, 137-152. https://doi.org/10.1016/j.jhydrol.2015.09.047

Zhang, X., Wang, J., Zwiers, F. W., & Groisman, P. Y. (2010). The influence of large-scale climate variability on winter maximum daily precipitation over North America. *Journal of Climate*, *23*(11), 2902-2915. https://doi.org/10.1175/2010JCLI3249.1

Zhang, Y., Ragettli, S., Molnar, P., Fink, O., & Peleg, N. (2022). Generalization of an Encoder-Decoder LSTM model for flood prediction in ungauged catchments. *Journal of Hydrology*, *614*, 128577. https://doi.org/10.1016/j.jhydrol.2022.128577

Zhang, Y., Wallace, J. M., & Battisti, D. S. (1997). ENSO-like interdecadal variability: 1900–93. *Journal of Climate*, *10*(5), 1004-1020. https://doi.org/10.1175/1520-0442(1997)010<1004:ELIV>2.0.CO;2

Zhang, Z., Zhang, Q., & Singh, V. P. (2018). Univariate streamflow forecasting using commonly used data-driven models: literature review and case study. *Hydrological Sciences Journal*, *63*(7), 1091-1111. https://doi.org/10.1080/02626667.2018.1469756

Zhao, C., & Brissette, F. (2022). Impacts of large-scale oscillations on climate variability over North America. *Climatic Change*, *173*(1), 1-21. https://doi.org/10.1007/s10584-022-03383-2

Zhao, Q., Liu, Z., Ye, B., Qin, Y., Wei, Z., & Fang, S. (2009). A snowmelt runoff forecasting model coupling WRF and DHSVM. *Hydrology and Earth System Sciences*, *13*(10), 1897-1906. https://doi.org/10.5194/hess-13-1897-2009

Zhao, T., Cai, X., & Yang, D. (2011). Effect of streamflow forecast uncertainty on real-time reservoir operation. *Advances in Water Resources*, *34*(4), 495-504. https://doi.org/10.1016/j.advwatres.2011.01.004

Zhao, T., Yang, D., Cai, X., Zhao, J., & Wang, H. (2012). Identifying effective forecast horizon for real-time reservoir operation under a limited inflow forecast. *Water Resources Research*, *48*(1). https://doi.org/10.1029/2011WR010623

Zhu, S., Luo, X., Yuan, X., & Xu, Z. (2020). An improved long short-term memory network for streamflow forecasting in the upper Yangtze River. *Stochastic Environmental Research and Risk Assessment*, *34*, 1313-1329. https://doi.org/10.1007/s00477-020-01766-4