

Deep Learning-Based Probabilistic Hierarchical Reconciliation for Hydrological and Water Resources Forecasting

by

Mohammad Sina Jahangir

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Civil Engineering

Waterloo, Ontario, Canada, 2024

© Mohammad Sina Jahangir 2024

Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

External Examiner	Hoshin Gupta Professor
Supervisor(s)	John Quilty Assistant Professor
Internal Member(s)	James Craig Professor Chul Min Yeum Assistant Professor
Internal-external Member	Andrea Brookfield Associate Professor

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

This thesis is partially the product of collaborative research, and co-authored publications as follows:

Chapter 2:

- Jahangir, M.S. and Quilty, J., 2022. Temporal hierarchical reconciliation for consistent water resources forecasting across multiple timescales: An application to precipitation forecasting. *Water Resources Research*, 58(6), p.e2021WR031862.

Chapter 3:

- Jahangir, M.S. and Quilty, J., 2024. Generative deep learning for probabilistic streamflow forecasting: Conditional variational auto-encoder. *Journal of Hydrology*, 629, p.130498.
- Jahangir, M.S., You, J. and Quilty, J., 2023. A quantile-based encoder-decoder framework for multi-step ahead runoff forecasting. *Journal of Hydrology*, 619, p.129269.

Chapter 4:

- Jahangir, M.S. and Quilty, J., 2024. Hierarchical deep learning for consistent multi-Timescale hydrological forecasting: An application to precipitation forecasting. *Water Resources Research*, Under review.

Abstract

Accurate, probabilistic, and consistent forecasts at different timescales (e.g., daily, weekly, monthly) are important for effective water resources management. Considering the different timescales together as a hierarchical structure, there is no guarantee that when forecast models are developed independently for each timescale in the hierarchy, they will result in consistent forecasts. For example, there is no guarantee that one-seven day(s) ahead forecasts from one model will sum to a weekly forecast from another model. Significant efforts have been made in the time-series forecasting community over the last two decades to solve this problem, resulting in the development of temporal hierarchical reconciliation (THR) methods. Until recently, THR methods had yet to be explored for hydrological and water resources forecasting. The main goal of this research is to introduce THR to the field of hydrological and water resources forecasting and to merge it with the latest advancements in deep learning (DL) to provide researchers and practitioners with a state-of-the-art model that can be used to produce accurate, probabilistic, and consistent multi-timescale forecasts. To achieve this goal, this research follows three interconnected objectives, each including a main contribution to the field of DL-based hydrological forecasting.

In the first main contribution of this research, the potential of THR to produce accurate and consistent hydrological forecasts was verified for the first time in hydrology through a large-scale precipitation forecasting experiment using 84 catchments across Canada. Three THR methods were coupled with three popular time-series forecasting models (exponential time-series smoothing, artificial neural network, and seasonal auto-regressive integrated moving average) for annual precipitation forecasting at monthly (12-steps ahead), bi-monthly (6-steps ahead), quarterly (4-steps ahead), 4-monthly (3-steps ahead), semi-annual (2-steps ahead), and annual (1-step ahead) timescales. It was confirmed that not only does utilizing THR guarantee forecast consistency across all timescales, but it can also improve forecast accuracy.

DL models are increasingly being used for hydrological modeling, particularly for lumped simulation, due to their ability to capture complex non-linear relationships within hydrological data as well as their efficiency in deployment. Likewise, the application of DL for hydrological forecasting has gained momentum recently. DL models can extract complex patterns from meteorological forcing data (e.g., precipitation) to forecast future streamflow, often leading to forecasts that are more accurate than current conceptual models. However, due to uncertainty in the phenomena affecting hydrological

processes, it is necessary to develop accurate probabilistic forecast models to provide insights for informed water management decisions.

In the second main contribution of this research, two novel state-of-art sequence-to-sequence probabilistic DL (PDL) models were developed, tested, and applied for short-term (one-seven day(s) ahead) streamflow forecasting in over 75 catchments with varied hydrometeorological properties across both the continental United States (CONUS) and Canada. The two designed models, namely quantile-based encoder-decoder and conditional variational auto-encoder (CVAE) showed superior performance compared to the benchmark long-short-term memory (LSTM) network considering forecast accuracy and reliability. Specifically, CVAE, a generative DL model that can estimate magnitudes of different sources of uncertainty (e.g., aleatoric, epistemic), proved to be effective in producing reliable forecasts for longer forecast lead times (three-seven days ahead).

Given the introduction of THR to the field of hydrological forecasting through the first main contribution, there is no guidance on how to couple THR with the latest advancements in DL, especially PDL, to produce accurate, and consistent probabilistic hydrological forecasts. Furthermore, existing methods for combining THR with DL models, particularly PDL models, suffer from several limitations. Firstly, almost all approaches treat THR as a post-processing step. Secondly, existing THR methods often lack adaptability, meaning they are unable to adjust properly to changing data distributions or new information. Finally, there is limited research on implementing probabilistic THR, a crucial aspect for making probabilistic forecasts consistent.

As the third main contribution, a hierarchical DL model (HDL) was introduced where THR was integrated directly into the DL model. Specifically, a custom THR layer was developed that can be combined with any DL model, much like a LSTM layer or a linear layer, to produce the proposed HDL. This integrated approach (via the new THR layer) allows any DL model to leverage temporal information across multiple timescales during training, perform probabilistic THR, and be efficient for real-time application. Furthermore, the proposed HDL is based on auto-regressive normalizing flows, a state-of-the-art generative DL model that is more flexible than CVAE in that it can non-parametrically estimate the probability distribution of the target variable (e.g., streamflow). HDL was tested on more than 400 catchments across CONUS for weekly streamflow forecasting at daily (seven-steps ahead) and weekly (one-step ahead) timescales. The performance of HDL was benchmarked against LSTM variants. HDL produced forecasts that had substantially higher accuracy than the LSTM variants and

simultaneously generated consistent forecasts at both daily and weekly timescales, without the need for post-processing (as in the vast majority of THR methods). The implementation of THR as a neural network layer allows it to be seamlessly combined with other DL layers. For example, the new THR layer can be coupled with physics-based differentiable routing layers for multi-timescale distributed hydrological forecasting.

It is expected that HDL will serve as a benchmark upon which future THR methods will be compared for streamflow forecasting. Furthermore, given the generality of the approach, HDL can be used for forecasting other important variables within hydrology (e.g., soil moisture) and water resources (e.g., urban water demand), as well as other disciplines, such as renewable energy (e.g., solar power).

Acknowledgements

I am particularly grateful to my supervisor, Prof. John Quilty, for his guidance and support throughout my Ph.D. His expertise in AI model development was invaluable. I especially appreciate his kind and understanding nature, which fostered a positive and productive environment.

I would like to thank my defense committee members, Prof. Hoshin Gupta from the University of Arizona, Prof. James Craig, Prof. Andrea Brookfield, and Prof. Chul Min Yeum from the University of Waterloo.

During my Ph.D., I received a PhD Thesis Completion Award, two Dr. T.E. Unny Memorial Awards, and four University of Waterloo Graduate Excellence Scholarships, which are hereby acknowledged.

I would like to thank my friends in Waterloo for their unwavering support and encouragement.

Finally, I embark on this next chapter of my journey feeling a sense of accomplishment and pride. While I may not have achieved everything, I set out to do, I can confidently say that I gave it my all. My perspective remains unchanged by these past four years: “Give me a place where I can stand – and I shall move the world”.

Dedication

To my spouse, whose love, patience, and tireless efforts made this journey possible.

To my parents, for their unwavering support and the foundation they laid for my success.

To my sister, whom I missed dearly throughout these years.

In the loving memory of my grandparents, whose guidance and love continue to inspire me.

Table of Contents

Examining Committee Membership	ii
Author’s Declaration.....	iii
Statement of Contributions	iv
Abstract.....	v
Acknowledgements.....	viii
Dedication.....	ix
List of Figures.....	xiv
List of Tables	xvii
List of Abbreviations and Acronyms	xviii
Chapter 1 Introduction	1
1.1 Problem Statement	1
1.2 Research Outline	5
1.3 Thesis Structure and Research Contributions.....	7
Chapter 2 Temporal Hierarchical Reconciliation for Consistent Water Resources Forecasting Across Multiple Timescales: An Application to Precipitation Forecasting	11
2.1 Introduction	12
2.1.1 Contribution and Outline.....	15
2.2 Temporal Hierarchical Reconciliation	16
2.3 Study Data	18
2.4 Methodology	20
2.4.1 Forecast Methods	20
2.4.2 Temporal Hierarchical Reconciliation Framework	22

2.4.3 Performance Measures	25
2.5 Result and Discussion	26
2.5.1 ORF Accuracy	26
2.5.2 Temporal Hierarchical Reconciliation	30
2.5.3 Temporal Hierarchy's Structure and Its Impact on Reconciliation	40
2.6 Summary, Conclusions, and Future Work	40
Chapter 3 Generative Deep Learning for Probabilistic Streamflow Forecasting: Conditional Variational Auto-Encoder	43
3.1 Introduction	44
3.1.1 Generative Deep Learning.....	47
3.1.2 Contributions and Outline	49
3.2 Dataset.....	50
3.3 Methods.....	52
3.3.1 Long Short-Term Memory Network	53
3.3.2 Quantile-Based Encoder-Decoder	55
3.3.3 Conditional Variational Auto-Encoder.....	55
3.3.4 Quantile-Based Conditional Variational Auto-Encoder (QCVAE)	56
3.3.5 Model Development.....	56
3.3.6 Forecast Evaluation	58
3.4 Results and Discussion.....	60
3.4.1 Evaluating Deterministic Performance	60
3.4.2 Evaluating Probabilistic Performance	63
3.4.3 Impact of Different Features of the Input Variables on Model Performance	64
3.5 Conclusion, Limitations, and Future Work	67

Chapter 4 Hierarchical Deep Learning for Consistent Multi-Timescale Hydrological Forecasting..	69
4.1 Introduction	70
4.1.1 Contribution and Outline.....	73
4.2 Related Work.....	74
4.2.1 Machine Learning-Based Reconciliation	74
4.2.2 Flow-Based Generative Deep Learning	76
4.3 Data and Case Study	77
4.4 Methodology	78
4.4.1 Temporal Hierarchical Reconciliation (THR).....	78
4.4.2 Normalizing Flow (NF) for Streamflow Forecasting.....	82
4.4.3 Proposed Model.....	83
4.4.4 Performance Assessment.....	86
4.5 Results and Discussion.....	86
4.5.1 Non-Reconciled Daily Forecasts.....	86
4.5.2 Non-Reconciled Weekly Results.....	87
4.5.3 Reconciled Forecasts using Hierarchical Deep Learning (HDL).....	89
4.6 Conclusions, Limitations, and Future Work	92
Chapter 5 Conclusions, Limitations, and Future Work	94
5.1 Conclusions	95
5.2 Limitations and Future Work	98
References.....	100
Appendix A Temporal Hierarchical Reconciliation for Consistent Water Resources Forecasting Across Multiple Timescales: An Application to Precipitation Forecasting (Chapter 2)	124

Appendix B Generative deep learning for probabilistic streamflow forecasting: conditional variational auto-encoder (Chapter 3).....	135
Appendix C Hierarchical Deep Learning for Consistent Multi-Timescale Hydrological Forecasting (Chapter 4).....	142

List of Figures

Figure 1-1. A three-level/timescale hierarchy example. Each node is associated with a forecast step.	2
Figure 1-2. General workflow of this research and the inter-connections among the chapters.	8
Figure 2-1. All (698) basins in the Canadian model parameter experiment database along with outlets of the 84 selected basins. Black triangles illustrate the outlets of the selected basins. The basins are divided into five classes based on monthly median precipitation (mm) for improved illustration.....	19
Figure 2-2. Temporal hierarchy for monthly data using the same index (i) for all aggregation levels. $i = 1, 2, \dots, T/12$, where T , is the number of observations. It should be noted that the structure below is only one of the possible presentations for aggregating monthly data with a yearly forecast horizon. This figure is adopted from Athanasopoulos et al. (2017).	23
Figure 2-3. Residual plot of original forecasts for the three different models: (a) exponential time-series smoothing (b) artificial neural network, and (c) seasonal auto-regressive integrated moving average at the monthly scale (i.e., 12-steps ahead forecasts) from 2005 to the end of 2010 for 84 selected basins. Each gray line represents the residuals (of the forecasts) for a specific station. The average error across all basins is shown with a bold solid line.	27
Figure 2-4. The non-parametric fitted probability density function (pdf) of the Nash-Sutcliffe efficiency values associated with the exponential time-series smoothing model at each timescale. The area under the dashed line (kernel density estimate) represents the area under the pdf curve (probability).....	28
Figure 2-5. The net percentage of improvement (based on the number of forecasts) due to different temporal hierarchical reconciliation methods (WLS _s , WLS _v , OLS, and BU) with respect to the original forecasts (ORF) based on DM-1 and DM-2 tests. Note: BU is not reported for the monthly scale as it is equal to the ORF.	32
Figure 2-6. The original forecasts (ORFs), observed data (OBS), WLS _s , and BU forecasts for two cases associated with the 4-monthly scale: first row figures (a, b, c) are associated with the best ORF based on the nRMSE (Canadian model parameter experiment [CANOPEX] IDs: 515, 25, and 25, respectively), while the second-row figures (d, e, f) are associated with “mediocre” forecasts (CANOPEX IDs: 422, 105, and 328, respectively).....	34

Figure 2-7. Violin plots representing the dispersion of relative ($[\text{reconciled-ORF}]/\text{ORF} \times 100\%$) nRMSE values of the temporal hierarchical reconciliation methods compared to original forecasts (ORFs) for monthly (first column), 2-monthly (second column), and 4-monthly (third column) timescales. Bottom-Up (BU) is equal to ORF for the monthly scale. The circle marker and the solid line show the mean and the two standard deviation range ($\text{mean} \pm 2 \times \text{standard deviation}$), respectively. The results are associated with filtered ORFs.	36
Figure 2-8. The relative normalized root mean square error (nRMSE) change of the temporal hierarchical reconciliation methods compared to original forecasts (ORFs) at different timescales ($[\text{reconciled-ORF}]/\text{ORF} \times 100\%$). The results are associated with ORFs having a Nash-Sutcliffe efficiency value greater than zero at the monthly scale. The x-axis shows the accumulated percentage of sorted monthly forecasts (based on the nRMSE). Note: that bottom-up is equal to ORF at the monthly scale resulting in no relative change.....	39
Figure 3-1. The 75 selected basins across Canada.	51
Figure 3-2. General ED, CVAE, and LSTM-DN structure scheme. $\mathbf{X}$ is the input array to the models, while $\mathbf{y}$ indicates the target (streamflow).	53
Figure 3-3. KGE violin plot for all lead times. The overall median (across all lead times) and median values for each lead time (#LT) are mentioned above each subplot. The dashed line indicates the desired KGE value (one).....	62
Figure 3-4. Comparison of average forecast KGE with simulation KGE.	63
Figure 3-5. PICP violin plot for all lead times. The overall median (across all lead times) and (#LT) median values for each lead time are mentioned above each subplot. The dashed line indicates the desired PICP value (90 %).	64
Figure 3-6. Impact of different features of the input variables on the performance of CVAEQ. KGE and PICP values are averaged across forecast lead times of one-seven days ahead. The fitted lines are based on linear regression where their associated coefficient (slope coefficient) and the associated p-values are shown in the boxes.	66
Figure 3-7. KGE vs. PICP for different LTs associated with CVAEQ. The slope coefficients and their associated p-values are shown in the box.....	67
Figure 4-1. The defined hierarchy. The lower level is associated with the daily scale, while the upper level is associated with the weekly scale.....	79

Figure 4-2. Hierarchical model for daily and weekly streamflow forecasting. $\mathcal{F}E$, $\mathcal{F}D$ indicate the number of features associated with ERA5 Land and Daymet, respectively. Similarly, ILE and ILE indicate the input lag (look-back period) associated with ERA5 Land and Daymet, respectively.	84
Figure 4-3. Stepwise comparison of the models' performance associated with the daily scale. The median reported at the top of each chart was calculated across all catchments and forecast steps.	88
Figure 4-4. Absolute and relative discrepancy between the two timescales.	89
Figure 4-5. Relative difference between reconciled and original forecasts at the weekly scale. The x-axis represents the threshold group (1-7), while the y-axis represents the relative change in different error metrics compared to the ORF.	90

List of Tables

Table 2-1. Basin-wise average of statistical parameters (mean, median, standard deviation, 95th Percentile, and fifth Percentile) associated with precipitation for the selected 84 basins at each timescale.....	19
Table 2-2. ORF analysis. Number of remaining datasets (from initial 84), the average NSE (-) value for each forecasting method after removing basins with NSE values (at the respective timescale) less than 0, and the average of NSE values across all basins	30
Table 2-3. Average nMAE (-) values for ORF and the temporal hierarchical reconciliation methods obtained by each model (ETS, ANN, and SARIMA). Based on the definition of BU, the monthly scale values listed for ORF and BU are equal. At each timescale, the results are associated with ORFs having an NSE greater than 0.....	37
Table 3-1. Statistical parameters of the meteorological variables and streamflow (calculated across the 75 basins).....	51
Table 3-2. Mean and std for each metric averaged across forecast lead times and all 75 basins compared to the mean and std of the simulations (averaged across all 75 basins).....	60
Table 4-1. Hyperparameters and their associated search space (see Figure 4-2).	85
Table 4-2. DM test percentage (%) of significant change in forecast accuracy at the daily level/timescale across 402 catchments compared to BU (ORF). The values in bold show higher percentages.	92
Table 4-3. DM test percentage (%) of significant change in forecast accuracy at the weekly level/timescale across 402 catchments compared to BU. The values in bold show higher probabilities	92

List of Abbreviations and Acronyms

Akaike information criterion	AIC
Artificial intelligence	AI
Artificial neural network	ANN
Bottom-up	BU
Canadian model parameter experiment	CANOPEX
Catchment Attributes and MEteorology for Large-sample Studies	CAMELS
Conditional variational auto-encoder	CVAE
Continental United States	CONUS
Deep learning	DL
Dense network	DN
Deterministic	DT
Diebold-Mariano	DM
Encoder-decoder	ED
Evidence lower bound	ELBO
Exponential time-series smoothing	ETS
Generative deep learning	GDL
Hierarchical deep learning	HDL
Hierarchical reconciliation	HR
Hydrologic unit	HU
Kling-Gupta efficiency	KGE

Kwiatkowski–Phillips–Schmidt–Shin	KPSS
Lead time	LT
Long short-term memory with a dense network	LSTM-DN
Long-short-term memory	LSTM
Look-back	LB
Machine learning	ML
Masked autoencoder for distribution estimation	MADE
Masked autoregressive flow	MAF
Mean absolute error	MAE
Mean prediction interval width	MPIW
Mixture Gaussian	MG
Monte-Carlo	MC
Multi-layer perceptron	MLP
Multiple timescale; Multi-timescale	MTS
Multi-timescale long short-term memory	MTS-LSTM
Nash-Sutcliffe efficiency	NSE
Normalized mean absolute error	nMAE
Normalized root mean square error	nRMSE
Normalizing flow	NF
Observed data	OBS
Ordinary least squares	OLS
Original forecast	ORF
Prediction interval coverage probability	PICP

Probabilistic deep learning	PDL
Probabilistic temporal hierarchical reconciliation	PTHR
Quantile-based conditional variational auto-encoder	QCVAE
Real-valued non-volume preserving method	Real NVP
Root mean squared error	RMSE
Seasonal and trend decomposition using loess regression	STL
Seasonal auto-regressive integrated moving average	SARIMA
Temporal hierarchical reconciliation	THR
Top-down	TD
U.S. Geological Survey	USGS
Variational auto-encoder	VAE
Weighted least squares	WLS

Chapter 1

Introduction

1.1 Problem Statement

Accurate forecasts of key hydrological and water resources variables (e.g., precipitation, streamflow, evaporation) across multiple timescales (MTS; e.g., timescales such as daily, monthly, yearly) are integral for the successful planning, management, and operation of water resources systems (van den Hurk et al., 2016). However, it is often the case that different models are applied for forecasting different timescales of a particular variable (Peters-Lidard et al., 2019), without ensuring that the forecasts are consistent across timescales (e.g., monthly forecasts aggregate to seasonal or yearly forecasts), which is also known as coherency (Taieb et al., 2017). In such cases, decision-makers are provided with disparate information that may lead to sub-optimal planning, management, and operation of water resources systems. Further, bottom-up (BU) forecasts, which aggregate finer scale forecasts to a coarser scale (see, for example Wang et al., 2011), miss an opportunity to exploit frequency characteristics of the data that may prove helpful in improving forecast quality (Spiliotis et al., 2020). In the BU method, as the most disaggregated (base) time-series are being used for building the coarser forecasts, obtaining accurate base forecasts is very important. High-frequency (lower-level series) could be noisy and hence, more challenging to forecast (Athanasopoulos et al., 2009). Therefore, the BU method may result in error propagation (through the higher-level series). Further, when using the BU method, the benefits of potentially more accurate forecasts at the disaggregated or aggregated levels cannot be leveraged through temporal/cross-temporal weighting (Yang et al., 2017). In general, hierarchical reconciliation (HR) methods have been developed to solve the inconsistency between the forecasts at different levels (Hyndman et al., 2011; Hyndman et al., 2016) while leveraging information from all levels. HR has been shown to outperform the BU method (Hyndman et al., 2011, 2016; Athanasopoulos et al., 2017; Kourentzes et al., 2017). In general, HR is a statistical method used to ensure that forecasts produced from different models, each at a specific level, are consistent with the forecasts at all other levels. HR applies to any time-series variable that can be aggregated across levels (e.g., precipitation, streamflow, evaporation). Thus, HR has potential for numerous hydrological and water resources applications (reservoir operation, irrigation scheduling, water supply, etc.).

A specific type of HR is temporal HR (THR), where forecasting at MTS (e.g., daily, weekly) is modeled as a hierarchy (timescales as hierarchy levels). Although not explicitly related to the hydrology and water resources domains, several studies have emphasized the benefits and limitations of temporal

aggregation (using multiple temporally aggregated series, similar, but different than THR) in improving forecast accuracy at different timescales (Athanasopoulos et al., 2017; Bergsteinsson et al., 2021; Kourentzes et al., 2017; Rostami-Tabar et al., 2014, 2013). THR has been growing in popularity across several fields (e.g., solar power, electricity demand). In this thesis, the term ‘reconciliation’ is used since forecasts are constructed for each timescale (hierarchy level, see below), either simultaneously or individually, and then combined to ensure the resulting forecasts are consistent. Unlike BU, where only the disaggregated series are forecasted, forecasts are produced at all timescales and reconciled afterward. Prior to this research, THR had yet to be explored in the field of water resources. An example of THR, is shown below in Figure 1-1, where at the most disaggregated timescale, hourly streamflow forecasts ($\hat{y}_{i=1-6}^1$) are produced, at the second timescale ($\hat{y}_{i=1-3}^2$), two-hourly forecast are produced, and at the most aggregated timescale, six-hourly forecast ($\hat{y}_{i=1}^3$), are produced. When the forecasts are obtained at each timescale independently, there is no guarantee the produced forecasts will be consistent. It is expected by implementing THR, not only consistent forecasts will be obtained, but the general accuracy across all timescales will be improved.

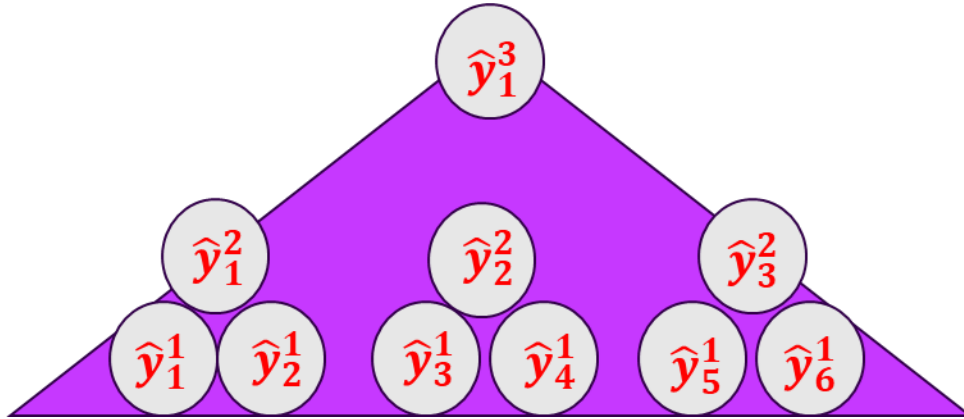


Figure 1-1. A three-level/timescale hierarchy example. Each node is associated with a forecast step.

To thoroughly evaluate the effectiveness of THR, it is crucial to first have high-quality forecast models. Using data-driven models, especially deep learning (DL) models, for hydrological prediction has been well-established (e.g., Kratzert et al., 2019, 2018; Yin et al., 2021; Zhang et al., 2021). Process-based models use linear and differential equations (Bittelli et al., 2010; Partington et al., 2012; Zhang et al., 2021) to describe the sub-processes driven by physical mechanisms (e.g., evapotranspiration, interception, infiltration; Xie et al., 2021). In contrast, data-driven models rely on data to learn the underlying function, and DL models are especially useful at this task when the data are plentiful

(Nearing et al., 2024). While process-based models provide better insight into physical processes than data-driven models (Zhang et al., 2015), they have two major drawbacks (Mehr et al., 2013; Ni et al., 2020). Firstly, because physical processes are often very complex, many simplifying assumptions are made during the modeling process. This results in uncertainties that are often ignored during model development and application (Jahangir et al., 2023; Moges et al., 2020). The second drawback of process-based models is their inefficiency and the considerable amount of information (i.e., types of data) they require as input. In contrast, data-driven models estimate statistical relationships between input/explanatory and target/response variables without considering the governing equations of the modeled system (Liu et al., 2015; Solomatine and Ostfeld, 2008). Because of this, data-driven models are often referred to as "black box" models. Compared to process-based models, data-driven models are more efficient in deployment and require less data (e.g., number of explanatory variables) during development (Jahangir et al., 2023). Previous studies on DL-based hydrological prediction have mainly focused on hydrological simulation (e.g., rainfall-streamflow modeling) using DL (e.g., Kratzert et al., 2018; Xie et al., 2021) instead of hydrological forecasting. The term forecasting is used in the machine learning (ML) domain when future values are predicted using information from the past and/or present (Jahangir et al., 2023; Quilty et al., 2023). Exploration of the design, optimization, and (hyperparameter-) tuning of DL models for hydrological forecasting has been under-explored in the literature.

Although previous research (e.g., Apaydin and Sibtain, 2021; Cheng et al., 2020; Cui et al., 2022; Han and Morrison, 2022; Kao et al., 2020; Xiang et al., 2020) has shown that DL models, especially those based on long short-term memory (LSTM) are effective for hydrological forecasting, there exists important aspects that have not yet been thoroughly investigated. A major drawback of the previous studies is their focus on deterministic forecasting. Due to the uncertainty associated with hydrological and water resources variables (precipitation, streamflow, etc.), it is desirable to develop probabilistic forecast models when working with such datasets. In other words, there is benefit to develop models that not only produce point forecasts of the variable of interest, but also indicate the uncertainty of the forecasts (using, for example, prediction intervals). Probabilistic DL (PDL) models, such as those introduced by Klotz et al. (2022), are lacking in the hydrological forecasting literature. Based on the type of uncertainty (e.g., input data, model), and the required computational efficiency, different PDL models can be used. For example, Bayesian DL leverages Bayes' theorem and can be used to estimate the posterior distribution of the weights and biases of a DL model (Luo et al., 2022), and is especially

suitable for quantifying epistemic (model) uncertainty (Li et al., 2021). However, the computational demands of exact Bayesian inference can be substantial, especially for complex (large) DL models. As an alternative, variational inference offers a more computationally efficient approach for approximating the posterior distribution of network weights or target distribution, often referred to as aleatoric uncertainty (see Blei et al., 2017; Klotz et al. 2022). While this method improves computational efficiency, the accuracy of the approximation can vary depending on the model complexity (Blei et al., 2017). Generative DL (GDL), a recent advancement in computer science, possesses the ability to learn the underlying data distribution (Dumas et al., 2022; Fanfarillo et al., 2021). While estimating the target (e.g., streamflow) distribution makes GDL particularly attractive for hydrological forecasting, the efficacy of GDL models' has yet to be thoroughly explored in the field of hydrology and water resources.

Another research avenue related to hydrological forecasting that requires further exploration is DL-based multi-timescale (MTS) hydrological forecasting. To date, no studies have specifically evaluated the use of a single DL model for either deterministic or probabilistic MTS hydrological forecasting. Although, two studies have implemented a single DL model for MTS hydrological simulation (Gauch et al., 2021; Wilbrand et al., 2023). These studies utilized a MTS LSTM model to predict both hourly and daily runoff simultaneously. While Gauch et al. (2021) demonstrated the efficacy of LSTMs for MTS runoff simulation, they sought to ensure the sub-daily predictions aggregated to the daily predictions (hereafter, referred to as consistency) by including a penalty in the loss/objective function. A key challenge in using a single model for MTS predictions is the potential for inconsistencies between predictions at different timescales. While regularizing the loss function can mitigate these inconsistencies, it does not eliminate them, as the inconsistencies remain unconstrained. To solve the inconsistency problem associated with MTS, THR could be adopted. On the other hand, for PDL-based MTS forecasting, probabilistic THR (PTHR; Jeon et al., 2019; Taieb et al., 2017; Yagli et al., 2019), a recent development in the time-series forecasting community, may be adopted. The effectiveness of THR and PTHR for improving the accuracy of hydrological forecasting is missing from the literature. Therefore, a thorough investigation is desirable to evaluate the potential of these two methods (THR and PTHR) combined with state-of-the-art DL models for MTS hydrological forecasting.

Previous studies outside of hydrology and water resources have shown success with THR (e.g., Athanasopoulos et al., 2017; Nystrup et al., 2020) and PTHR (e.g., Jeon et al., 2019; Yagli et al., 2020) in producing accurate and consistent MTS forecasts. However, these methods have limitations that

hinder their application in real-world scenarios, especially with the rapid advancements in artificial intelligence (AI)-based model development. One major drawback of previous studies focused on THR is their inability to adapt to new information. Due to their formulation, THR and PTHR rely solely on historical data, leading to static models that do not learn from new observations. This rigidity in their formulation limits their effectiveness in dynamic environments, such as those studied in hydrology, where conditions can change rapidly. Furthermore, their formulations might depend on pre-assumed structures within the data, which may not always hold true in complex hydrological systems. Another significant drawback is that THR (and PTHR) are used as post-processing techniques, which may be inefficient during the deployment phase in DL pipelines. In other words, they are applied after the initial forecast is made. This especially can be inefficient within DL pipelines, where seamless integration throughout the forecasting process is crucial for efficient and effective deployment.

1.2 Research Outline

This thesis develops a framework that can be used for producing accurate and consistent probabilistic forecasts across multiple timescales. The results are expected to be useful to hydrologists and water resources practitioners for tasks such as (but not limited to) flood forecasting, reservoir management, and optimizing water distribution system operations. The framework is developed in three stages by accomplishing three main objectives:

- i. Evaluating the potential of THR for hydrological and water resources forecasting using deterministic time-series models. THR is explored for the first time for hydrological and water resources forecasting and its added value compared to the traditional BU method is established using a precipitation forecasting experiment across monthly-annual timescales in Canada. The result is the first benchmark for future developments in THR-based hydrological and water resources forecasting.
- ii. Introducing GDL, a specific PDL variant, for accurate and reliable probabilistic hydrological forecasting. For this purpose, a novel GDL model is introduced to the field of hydrology and water resources. GDL is compared against the current state-of-the-art DL methods, such as LSTM, for probabilistic streamflow forecasting at the daily timescale in Canada and the continental United States (CONUS).
- iii. Designing and deploying a hierarchical DL (HDL) model for accurate and consistent MTS hydrological forecasting. The HDL model tackles the two associated problems with THR and

PTHR, which are the formulation being static (invariant to new observations) and being utilized as a post-processing tool. Furthermore, for the first time in the literature, PTHR is integrated within GDL and explored in detail. To assess the effectiveness of the HDL model, the model is compared to both deterministic and probabilistic DL benchmarks.

Water resource management relies heavily on accurate and consistent forecasts at different timescales (e.g., daily, weekly, monthly). However, developing independent models for each timescale can lead to inconsistencies, which can be misleading for decision-makers. This thesis addresses this challenge by exploring the application of three THR methods for hydrological and water resources forecasting. To demonstrate its effectiveness, this thesis investigates THR's applicability for monthly and annual precipitation forecasting across > 80 Canadian catchments. THR is coupled with three common statistical time-series forecasting models, and its performance is compared to the traditional BU aggregation, where monthly forecasts are simply summed to obtain annual forecasts.

DL models are increasingly utilized in hydrological modeling, especially for lumped simulations, due to their ability to capture complex non-linear relationships within hydrological data. This capability leads to more accurate streamflow simulations compared to other data-driven methods and conceptual models. Recently, DL applications in hydrological forecasting have gained momentum as these models effectively capture and process complex patterns in hydrometeorological forcing data to predict future streamflow. This provides valuable insights for informed water management decisions. Given the stochastic nature of phenomena affecting hydrological forecasting, developing accurate and reliable probabilistic forecast models is essential. In this thesis, two novel state-of-the-art PDL models are developed, deployed, and tested for streamflow forecasting on more than 75 catchments across CONUS and Canada, each with varied hydrometeorological properties. The two models, quantile-based encoder-decoder and the GDL model, conditional variational auto-encoder (CVAE), are not only able to produce accurate hydrological forecasts, but also generate reliable prediction intervals for the forecasts. By using the proposed GDL model, one can incorporate uncertainties from both input data and the model structure into the modeling process, leading to more robust forecasts.

Prior to this research, THR had yet to be integrated with DL for hydrological and water resource prediction. Existing methods for combining THR with DL models, especially PDL models, suffer from inefficiency due to post-processing and lack of adaptability to adjust to changing data distributions or new information. Also, there is limited research on implementing PTHR, a crucial aspect for combining

probabilistic forecasts. The main goal of this thesis is to present a novel approach to address these shortcomings. To tackle this goal, the HDL model is introduced, integrating THR directly into the DL model. This inline approach allows the DL model to leverage temporal information during training, perform PTHR, and be efficient for deployment. In the short-term, HDL is anticipated to serve as a benchmark for MTS hydrological forecasting in the research community and in the long-term, adopted by water resources practitioners. In addition to providing consistent and accurate forecasts, HDL's modular design offers the capability to update the model with the latest advancements in the AI community, ensuring its ongoing relevance and effectiveness.

1.3 Thesis Structure and Research Contributions

This thesis is organized around published and submitted manuscripts. Chapter 2 is based on a manuscript published in the Water Resources Research (Jahangir and Quilty, 2022). Chapter 3 closely aligns with two published papers in Journal of Hydrology (Jahangir et al., 2023; Jahangir and Quilty, 2024). Finally, Chapter 4 is based on a manuscript submitted to the journal, Water Resources Research. The general workflow of the research, and the inter-connections between thesis' chapters are illustrated in Figure 1-2.

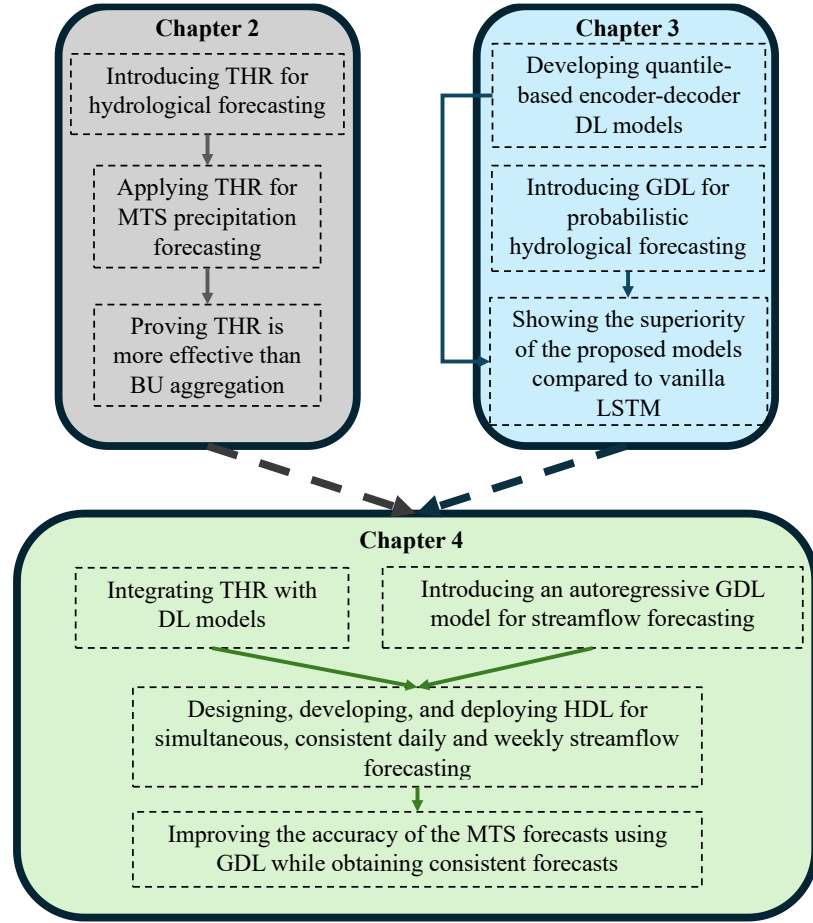


Figure 1-2. General workflow of this research and the inter-connections among the chapters.

This research has made the following main contributions to the DL-based hydrological and water resources forecasting literature:

- i. The first main contribution of this research demonstrated the potential of THR to generate accurate and consistent forecasts as validated through a large-scale precipitation forecasting experiment involving 84 catchments across Canada. Three THR methods were integrated with three popular time-series forecasting models, exponential smoothing, artificial neural networks, and seasonal autoregressive integrated moving average, to forecast annual precipitation at various timescales: monthly (12 steps ahead), bi-monthly (6 steps ahead), quarterly (4 steps ahead), 4-monthly (3 steps ahead), semi-annual (2 steps ahead), and annual (1 step ahead). The results confirmed that using THR not only ensures forecast consistency

across all timescales but also enhances forecast accuracy. Details regarding the case study, THR implementation, and results are provided in Chapter 2.

- ii. The second main contribution of this thesis is the development and application of advanced PDL models for short-term (daily) streamflow forecasting (one to seven days ahead). The thesis introduces two state-of-the-art sequence-to-sequence PDL models: the quantile-based encoder-decoder and the CVAE (also its quantile-based counterpart). These models were tested across over 75 catchments with diverse hydrometeorological properties in CONUS and Canada, demonstrating superior performance in forecast accuracy and reliability compared to traditional LSTM models. Notably, the CVAE model, capable of estimating different sources of uncertainty (i.e., both aleatoric and epistemic uncertainties), proved highly effective for longer lead times (three to seven days ahead) and challenging forecasting conditions (e.g., catchments with lower baseflow index, as well as those that are mainly precipitation driven). The findings underscored the potential of these novel PDL models for more accurate and reliable hydrological forecasting, paving the way for their adoption by researchers and water resource decision-makers. In Chapter 3, a comprehensive overview of the model development procedure, hyperparameter optimization, and evaluation of the forecasting models is provided.
- iii. The third main contribution of this thesis is the introduction of the HDL model that seamlessly integrates THR directly into a single DL model. This was achieved by developing a custom THR layer that can be incorporated into any DL model. This integrated approach enables the DL model to leverage temporal information across MTS during training, perform PTHR, and remain efficient for real-time applications. The proposed HDL model is based on autoregressive normalizing flows (NFs), a state-of-the-art GDL model that offers greater flexibility by non-parametrically estimating the probability distribution of the target variable (e.g., streamflow). The HDL model was tested on over 400 catchments across CONUS for weekly streamflow forecasting at both daily (seven-steps ahead) and weekly (one-step ahead) timescales. When benchmarked against LSTM variants, the HDL model demonstrated substantially higher forecast accuracy and generated consistent forecasts at both daily and weekly timescales without requiring post-processing, unlike most existing THR methods. An innovative feature of implementing THR as a custom layer not only allows it to be coupled with any type of DL model but also facilitates its integration with other DL layers. For example, the THR layer could be combined with physics-based layers, such as differentiable routing

layers, for MTS distributed hydrological forecasting. It is anticipated that the HDL model will become a benchmark for future THR methods in streamflow forecasting. Moreover, due to the generality of this approach, the HDL model can be applied to forecasting other important variables in hydrology (e.g., soil moisture) and water resources management (e.g., urban water demand), as well as in other fields such as renewable energy (e.g., solar power forecasting). Details about the HDL model development, deployment, and testing are provided in Chapter 4.

Chapter 2

Temporal Hierarchical Reconciliation for Consistent Water Resources Forecasting Across Multiple Timescales: An Application to Precipitation Forecasting

This chapter is largely based on an article of the same title published in the journal, *Water Resources Research*. Some adjustments have been made to the original text to ensure better integration with the thesis. As a result, while the essence remains, there may be slight differences between the manuscript and the thesis.

Jahangir, M.S. and Quilty, J., 2022. Temporal hierarchical reconciliation for consistent water resources forecasting across multiple timescales: An application to precipitation forecasting. *Water Resources Research*, 58(6), p.e2021WR031862.

Summary

Obtaining consistent forecasts at different timescales is important for reliable decision-making. This study introduces and evaluates the benefits of utilizing temporal hierarchical reconciliation methods for water resources forecasting, with an application to precipitation. Original (precipitation) forecasts (ORFs) were produced using “automatic” exponential time-series smoothing (ETS), artificial neural network (ANN), and seasonal auto-regressive integrated moving average (SARIMA) models at six timescales, namely, monthly, 2-monthly, quarterly, 4-monthly, bi-annual, and annual, for 84 basins extracted from the Canadian model parameter experiment. Temporal hierarchical reconciliation methods, including structural scaling-based weighted least squares (WLS), series variance scaling-based WLS, and ordinary least squares, along with the simple bottom-up (BU) method, were applied to reconcile the forecasts. In general, ETS (direct forecasting) demonstrated better performance compared to ANN and SARIMA (recursive forecasting). The results confirmed that improvements in accuracy due to reconciliation is dependent on the basin, timescale, and the ORFs’ accuracy. For different forecast models, the reconciliation methods showed different levels of performance. For ETS, BU was able to improve forecast accuracy to a greater extent than the temporal hierarchical reconciliation methods, while for ANN and SARIMA, forecast accuracy was improved through all temporal hierarchical reconciliation methods but not BU. The reconciled forecasts’ accuracy was affected more by the ORFs’ accuracy than by the reconciliation method. Different timescales showed

dissimilar sensitivity to reconciliation. The presented results are anticipated to serve as a valuable benchmark for evaluating future developments in the promising area of temporal hierarchical reconciliation for water resources forecasting.

2.1 Introduction

Accurate forecasts of key water resources variables (e.g., precipitation, streamflow, evaporation) across multiple timescales (MTS; e.g., daily, monthly, yearly) are integral for the successful planning, management, and operation of water resources systems (van den Hurk et al., 2016). However, in water resources, it is often the case that different models are applied for forecasting different timescales of a particular variable (Peters-Lidard et al., 2019), without ensuring that the forecasts are consistent across timescales (e.g., that monthly forecasts aggregate to seasonal or yearly forecasts), providing decision-makers with disparate information that may lead to sub-optimal planning, management, and operation of water resources systems. Further, “bottom-up” (BU) forecasts, which aggregate finer scale forecasts to a coarser scale (see, e.g., Wang et al., 2011), miss an opportunity to exploit low- and high-frequency characteristics of the data that may prove helpful in improving forecast quality (Spiliotis et al., 2020). To overcome these significant issues, that have been mostly ignored in the water resources forecasting literature, this chapter explores the use of temporal hierarchical reconciliation (THR; Athanasopoulos et al., 2017), a development in the time-series forecasting community, for generating consistent forecasts of water resources variables across MTS. Briefly, THR is a statistical procedure used to ensure that forecasts produced from different models, each at a specific timescale, are consistent with the forecasts at all other timescales (see Section 2.2 for more details). The approach explored here applies to any variable that can be aggregated across timescales (e.g., precipitation, evaporation, streamflow volumes) and thus, should prove useful in numerous water resources applications (reservoir operation, irrigation scheduling, water supply, etc.).

Since THR has yet to be explored within the water resources domain, a brief explanation is given on the differences between two common reconciliation methods and two broader approaches, more commonly adopted in the hydrology and water resources domains, when working with time-series at MTS (namely, temporal upscaling and disaggregation) before introducing the adopted case study and statistical methods used for assessing the proposed temporal reconciliation approaches.

For consistent forecasting at different timescales, forecasts at higher hierarchy levels (i.e., more coarsely aggregated time-series) must be equal to the sum of lower-level forecasts. The previously

mentioned bottom-up (BU) method is a common approach (Dangerfield and Morris, 1992) and is sometimes referred to as temporal upscaling in hydrological studies, where both direct (Wang et al., 2011) and indirect (where time-series are fed to a model; Ryu et al., 2012) implementations have been adopted. Another common reconciliation method is the top-down (TD) method (Dangerfield and Morris, 1992; Hyndman et al., 2011; Villegas and Pedregal, 2018), where the aggregated series are used for producing the lower-level series, resulting in consistent forecasts among all levels (Athanasopoulos et al., 2009). In the hydrology and water resources domains, a similar concept named temporal disaggregation exists (Kumar et al., 2000). However, due to the uncertain nature of hydrological forcings/variables (precipitation, evaporation, etc.), stochastic disaggregation methods (see, e.g., Koutsoyiannis et al., 2003; Lombardo et al., 2017; Tsoukalas et al., 2019) have been utilized, where preserving the joint probability distribution of multivariate time-series is sought when disaggregating the data. Both temporal upscaling and stochastic disaggregation methods, unlike reconciliation methods (where the main goal is to obtain consistent forecasts at different timescales), have been adopted for producing consistent ensemble synthetic data for hydrological models and obtaining data at sites without measurements at timescales of interest (Koutsoyiannis et al., 2003). Unlike previous research conducted in the water resources domain, this work focuses on utilizing different reconciliation methods for obtaining consistent precipitation forecasts at different timescales.

In the BU method, as the most disaggregated (base) time-series is being utilized for building the upper-level forecasts, obtaining accurate base forecasts is very important. High-frequency series (lower-level series) could be noisy and hence, more challenging to forecast (Athanasopoulos et al., 2009). Therefore, the BU method may result in error propagation (through the higher-level series). In TD method, the main drawback is the loss of information due to disaggregation (e.g., seasonality, Athanasopoulos et al., 2009). Further, it is noted that using the latter (TD or BU) methods, the benefits of potentially more accurate forecasts at the disaggregated or aggregated levels cannot be leveraged through temporal weighting (Yang et al., 2017). Therefore, the present study aims to leverage three other methods introduced by Hyndman et al. (2011) and Athanasopoulos et al. (2017) for THR as post-processing techniques to obtain consistent forecasts and evaluate whether the accuracy of statistical (data-driven) models can be improved through reconciliation. While data-driven models are adopted here for forecasting, the explored framework is equally applicable to process-based (physical and conceptual) models.

Precipitation forecasting is used in this study to explore THR, as precipitation is one of the most important variables in water resources systems. Thus, any advancement in the capability to accurately forecast precipitation may also result in improved: hydrological models (Goodarzi et al., 2019; Kobold and Sušelj, 2005; Li et al., 2019; Yao et al., 2019), climate change scenario analysis, coastal risk management (Ketabchi and Jahangir, 2021, 2019) reservoir and water supply management, flood risk evaluation, hydro-power generation forecasting, agricultural planning, and so on (Choubin et al., 2018; Field et al., 2014; Gouda et al., 2020). Of note, each of the previously mentioned applications requires accurate precipitation forecasts across different timescales.

While numerical models have been widely adopted for precipitation analysis, simulation, and forecasting (see, e.g., Dravitzki and McGregor, 2011; Goswami and Mohapatra, 2014; Jee and Kim, 2017; Li et al., 2016; Xuan et al., 2009), it has been shown that due to the uncertainties associated with physical parameterization schemes, precipitation forecasts obtained from numerical models are less reliable compared to directly resolved variables such as temperature and air pressure (Betts et al., 1998; Bukovsky and Karoly, 2007; Higgins et al., 1996; Pan et al., 2019; Tian et al., 2017; Vitart, 2004). Data-driven models, on the other hand, because of the recent increase in computing power and data availability as well as developments in parallel computing, have shown promising results in precipitation forecasting (Lu and Wang, 2011; Papacharalampous et al., 2018a; Sønderby et al., 2020; Wang et al., 2014; Yeditha et al., 2020). Furthermore, the significance of univariate time-series forecasting in the hydrometeorology and water resources domains has been well documented (the interested reader is directed to Green et al. (2009) and Papacharalampous et al. (2019) more details regarding this matter). Therefore, in the present study, univariate time-series forecasting using data-driven methods is chosen for precipitation forecasting (please see Section 2.4). Hereafter, the term “forecast” is referred to as data-driven point forecasting.

One of many possible ways to classify forecast methods is to divide forecast methods into two groups: automatic and non-automatic. Non-automatic methods refer to forecasting problems requiring careful user intervention for each case, whereas automatic forecasting is used when no or minimal user intervention is made during the forecast process (Chatfield, 1988). Automatic forecasting is vital when many time-series are required to be forecasted (Papacharalampous et al., 2018b), and user intervention in each case is not possible or highly time- or cost-prohibitive. Examples of automatic forecasting tools include “forecast” (Hyndman and Khandakar, 2008) and “prophet” (Taylor and Letham, 2018) packages. Papacharalampous et al. (2018a) evaluated monthly air temperature and precipitation

forecasting predictability using these two packages through naïve, random walk (Hyndman and Athanasopoulos, 2018), auto-regressive integrated moving average (ARIMA; Valipour et al., 2013), exponential time-series smoothing (ETS; Hyndman et al., 2008), ETS state-space model with Box-Cox transformation, auto-regressive moving average errors model with trend and seasonal components (De Livera et al., 2011), Theta (Assimakopoulos and Nikolopoulos, 2000), and Prophet (Taylor and Letham, 2018) methods. Their results suggested that all considered methods are suitable for long-term (monthly) air temperature and precipitation forecasting apart from naïve and random walk methods. In the present study, ETS, artificial neural network (ANN, please refer to Zhang et al. (1998), and Zhang (2001) for detailed review on implementing ANN for time-series forecasting), and seasonal auto-regressive integrated moving average (SARIMA) methods from the “forecast” package (Hyndman et al., 2022; Hyndman and Khandakar, 2008) are applied for multi-step ahead (up to 1 year ahead) precipitation forecasting at six different timescales (monthly, 2-monthly, quarterly, 4-monthly, bi-annual, and annual) across 84 basins in Canada (see Section 2.3).

2.1.1 Contribution and Outline

This study’s main objective is to implement THR combined with different data-driven forecasting methods, at multiple sites, for obtaining consistent forecasts at different timescales. Exploring temporal aggregations across MTS can reveal important information about the underlying process (Nystrup et al., 2020). However, when forecasts are independently produced (through different models) at different timescales, they usually do not result in consistent forecasts across a particular horizon (e.g., aggregating 1–3 months ahead forecasts do not sum to the seasonal forecast). This property is usually referred to as incoherency in the literature (N. Kourentzes and Athanasopoulos, 2019). This study explores how implementing THR can be leveraged as a post-processing technique to obtain consistent forecasts and improve the forecasts’ accuracy at different timescales, focusing on precipitation but with broad application to many water resources variables. The goal of such an approach is to generate consistent forecasts across MTS that can support the operation, management, and planning of water resources systems, helping to avoid incoherent or contradictory decisions (Nystrup et al., 2020; Yagli et al., 2019). To the authors’ best knowledge, no previous work has examined how THR impacts water resources forecasting accuracy at different timescales, let alone using a large-scale (multi-site) case study. This represents the first contribution of this study. As a second contribution, this study also evaluates the selected forecasting methods’ (ETS, ANN, and SARIMA) effect on the temporally reconciled results. While the initial assessment of the proposed method focuses on precipitation, it is

believed that its application to evapotranspiration, streamflow, and water quality parameters such as phosphorus and chlorophyll-a could be fruitful. The third contribution of this study assesses the impacts that the original forecasts (ORFs) and reconciliation methods' structure (such as the error variance matrix structure, see Section 2.4) have on the final reconciled forecasts' accuracy. Finally, this study's results could be used as a benchmark for other related studies aiming to construct consistent precipitation forecasts at multiple sites through THR.

The rest of the chapter is structured as follows: background on THR is given in Section 2.2, the data and adopted pre-processing methods are discussed in Section 2.3, the methodology is explained in detail in Section 2.4, the results are presented and analyzed in Section 2.5, and the chapter is concluded, in Section 2.6.

2.2 Temporal Hierarchical Reconciliation

Although not explicitly related to the water resources domain, several works have emphasized the benefits and limitations of temporal aggregation (utilizing different temporal aggregated series) on improving forecast accuracy at different timescales (e.g., Abraham, 1982; Athanasopoulos et al., 2017, 2011; Kourentzes and Petropoulos, 2016; Petropoulos and Kourentzes, 2015; Rostami-Tabar et al., 2014, 2013; Silvestrini and Veredas, 2008 among many others). It has been shown that, generally, the higher levels of the temporal hierarchy (i.e., finer resolutions) demonstrate improved forecast accuracy after reconciliation (Athanasopoulos et al., 2017; Rostami-Tabar et al., 2013). Previous studies have also highlighted the significance of auto-correlation (Tiao, 1972) and stationarity (Rostami-Tabar et al., 2014) of the time-series (disaggregated series) with respect to improvements in forecast accuracy gained through temporal aggregation, where stationarity of the original time-series resulted in improvements in accuracy while highly auto-correlated time-series showed less improvement through aggregation. Different approaches exist for producing consistent forecasts at different timescales. Among these, the two most common methods are BU and TD methods (see Section 2.1). Another group of temporal aggregation methods that have recently gained popularity, and the focus of this study, are THR methods. The term "reconciliation" is used since forecasts are constructed for each timescale (hierarchy level) and then combined to ensure the resulting forecasts are consistent. Unlike BU and TD methods, where only the most aggregated and disaggregated series are being forecasted, respectively, forecasts are produced at all levels and reconciled afterward.

The forecasts produced at each level (before reconciliation is applied) are called base forecasts or ORFs. Previous studies have shown that reconciliation could improve the ORFs' accuracy along with constructing consistent forecasts. Hyndman et al. (2011) formulated the forecast reconciliation problem as a linear regression problem. Instead of estimating the ORFs' covariance (see Hyndman et al., 2011 and Section 2.4 for details), the authors used the ordinary least squares (OLS) method for estimating the regression coefficients. Later, Hyndman et al. (2016) proposed weighted least squares (WLS) for calculating the regression coefficients. Their method only considered the diagonal entries of the covariance matrix (i.e., the variances). Athanasopoulos et al. (2017) considered three diagonal estimators as approximations of the error sample covariance matrix, showing that the estimation method solely relying on the hierarchy structure could outperform other methods (e.g., BU, OLS). Wickramasuriya et al. (2019) extended the previous authors' work by considering the off-diagonal entries of the base forecast error covariance matrix. Nystrup et al. (2020) proposed a method for estimating autocorrelation within each aggregation level to improve reconciliation accuracy. Their results indicated that the reconciled forecasts at all timescales improve by considering the autocorrelation and cross-correlation of the ORFs.

The present work adopts the BU and OLS methods and two methods proposed by Athanasopoulos et al. (2017) for THR, structural scaling-based WLS (WLS_S), where the covariance error matrix weights are proportional to the hierarchy structure, and series variance scaling-based WLS (WLS_V), where the covariance error matrix weights are associated with the series in-sample error variance (see Section 2.4). The proposed methods' effectiveness was also confirmed by Yang et al., (2017) for solar power forecasting and Nystrup et al. (2020) for electricity load forecasting. However, the methods have not yet been assessed for water resources forecasting. The forecasts' timescales of the present study are similar to Athanasopoulos et al. (2017) and include monthly up to annual scales, which is temporally coarser compared to time-series used in Yang et al. (2017) and Nystrup et al. (2020) (hourly to daily). In general, and for water management purposes, monthly forecasting is important in managing dammed reservoirs, whereas daily forecasting is usually more relevant to the management of urban reservoirs (Banihabib and Mousavi-Mirkalaei, 2019). Although more recent methods exist for THR (such as probabilistic THR, see Athanasopoulos et al., 2020 and Yagli et al., 2019), they are out of the scope of the present work as this research focuses on (a) introducing the concept of THR for water resources forecasting and (b) applying it to a large-scale (multi-site) precipitation forecasting case study. There has been a recent, but inconclusive, debate over identifying the "best" THR method; however, it has

been shown that the results are case study and timescale-dependent (Villegas and Pedregal, 2018). The latter highlights the importance of exploring different THR methods for water resources forecasting. In the next section, the case study data are discussed before describing the methodology in detail. To distinguish between the methods leveraging forecasts at different levels of the temporal hierarchy (WLS_s, WLS_v, OLS) and BU (which only uses forecasts at the most disaggregated level), the term “THR” is used for WLS_s, WLS_v, and OLS throughout this chapter.

2.3 Study Data

Average daily total precipitation (measured in mm) for 698 catchments is available in the Canadian model parameter experiment (CANOPEX) database (Arsenault et al., 2016). The data contains two different sources of basin-averaged meteorological data: weather stations and gridded climate products. In this study, the data provided by Natural Resources Canada (daily 10 km gridded data, please see Hutchinson et al., 2009 for more details) is used as it does not contain any missing data (for more details regarding the Canadian model parameter experiment [CANOPEX] database, please see Arsenault et al., 2016). From all available basins, only 84 basins covering the period of 1950–2010 (the most extensive period available in the dataset), are chosen for constructing the ORFs (i.e., the base forecasting models). Daily data are aggregated at monthly, 2-monthly, quarterly (seasonal), 4-monthly, bi-annual, and annual timescales (please see Section 2.4). Forecasts are produced using the precipitation time-series from each of the 84 basins. The data is divided into two partitions, training (1950–2004) and testing (2005–2010). The forecast horizon is set to annual, resulting in a 12-steps, 6-steps, 4-steps, 3-steps, 2-steps, and a 1-step ahead forecast at the monthly, 2-monthly, quarterly, 4-monthly, bi-annual, and annual timescales, respectively. All (698) basins were grouped into one of five classes (according to their median monthly precipitation), and the 84 basins adopted in this study (indicated by a point at the basins' outlet) are illustrated in Figure 2-1. Figure 2-1 shows that the 84 selected basins are mainly located in southern Canada, with a higher relative density in the west compared to central and eastern regions. It is also shown that all five classes (monthly median precipitation ranging from 15 to 160 mm) are included among the 84 basins used in the model building procedure.

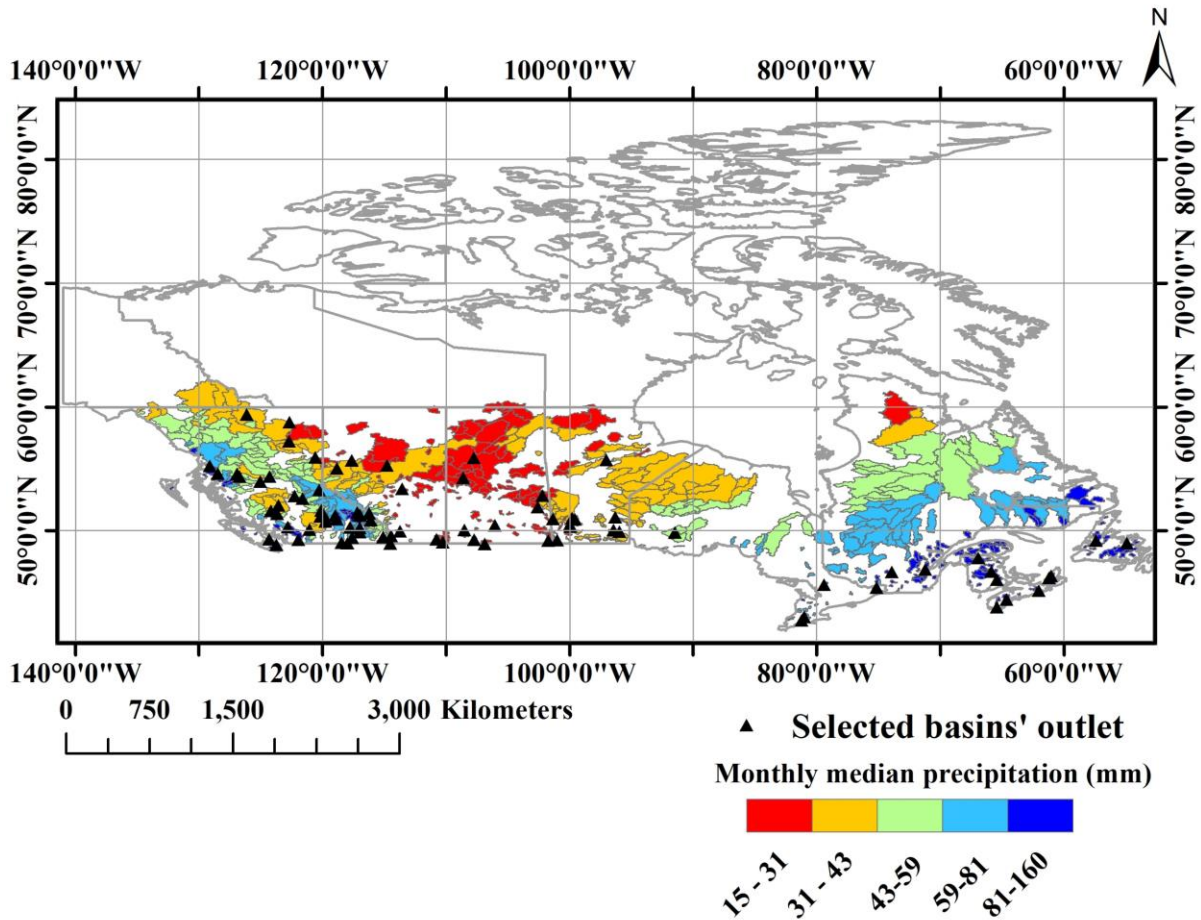


Figure 2-1. All (698) basins in the Canadian model parameter experiment database along with outlets of the 84 selected basins. Black triangles illustrate the outlets of the selected basins. The basins are divided into five classes based on monthly median precipitation (mm) for improved illustration.

The statistical parameters mean, median, standard deviation, 95th percentile, and fifth percentile of precipitation data, averaged across all 84 basins, for each of the six timescales, are listed in Table 2-1. The values are associated with the training section of data (1950–2004; the statistical parameters of the test set [2005–2010] are provided in Appendix A). Based on the diversity (location) of the selected basins (please see Figure 2-1) and the variation of precipitation among the basins (see Table 2-1), the results of the present study could be used as a benchmark for future studies exploring different temporal hierarchical reconciliation methods for consistent annual precipitation forecasting.

Table 2-1. Basin-wise average of statistical parameters (mean, median, standard deviation, 95th Percentile, and fifth Percentile) associated with precipitation for the selected 84 basins at each timescale.

Timescale	Statistical parameter (mm)				
	Mean	Median	Standard Deviation	95 th percentile	5 th percentile
Monthly	66.08	59.25	36.40	135.24	19.71
2-Monthly	132.17	122.46	57.44	239.93	56.05
Quarterly	198.25	186.31	75.63	333.44	95.69
4-Monthly	264.34	252.62	89.41	422.67	139.98
Bi-annual	396.51	393.17	85.25	538.27	268.22
Annual	793.01	793.35	120.64	980.47	610.47

2.4 Methodology

This section describes the forecasting methods used for obtaining the ORFs, followed by a detailed explanation of the THR framework and error metrics adopted for evaluating the performance of the different forecasts.

2.4.1 Forecast Methods

2.4.1.1 ETS

ETS models can be divided into 30 sub-types based on the method of representing the trend component (none, additive, additive-damped, multiplicative, multiplicative-damped), the seasonal component (none, additive, multiplicative), and the error component (additive, multiplicative) of a given time-series (Hyndman et al., 2008, 2022; Hyndman and Khandakar, 2008). However, some ETS combinations (such as multiplicative error combinations) may result in numerical instability (Hyndman & Khandakar, 2008) and should be avoided for forecasting. In this study, the models' hyperparameters (order of differencing, number of lags, etc.) were selected based on the minimum obtained Akaike information criterion (AIC; Akaike, 1974) at each timescale. The interested reader may refer to (Hyndman et al., 2008, 2022) for more details on ETS forecasting. Model selection, model fitting, and forecasting were carried out through custom scripts in R and the “forecast” package (Hyndman et al., 2022; Hyndman & Khandakar, 2008).

2.4.1.2 ANN

ANN models attempt to imitate the data processing of the human brain and the nervous system. These models have been proven to serve as a universal approximator when provided with sufficient degrees of freedom (Hornik et al., 1989). Neural networks and their variants (such as generalized regression neural networks, see Yan, 2012), especially recurrent neural networks (Hewamalage et al., 2021), have been shown to be very useful for time-series forecasting. In the present study, a multi-layer feed-forward backpropagation ANN (Abhishek et al., 2012) is adopted where the first layer serves as the input layer, the second layer (hidden layer) is the transformation layer, and the third layer is the output layer. The input layer consists of the seasonal and non-seasonal lags of the precipitation data. For each basin, and at each timescale, 20 models were trained (using different initial weights of the network, a default setting for the adopted software package) and then averaged for forecasting. The number of inputs (specific lags to be used in the network) was selected based on the AIC value. The ANN can be mathematically expressed as (Hyndman and Athanasopoulos, 2018):

$$y_{t+h} = f(\mathbf{U}_{(t+h)-1}) + \varepsilon_{t+h} \quad \text{Eq 2-1}$$

where h is the forecast horizon, f is the feed-forward neural network, $\mathbf{U}_i$ is the input vector containing all the selected lags (seasonal and non-seasonal) up to the i th step, and ε_i is a homoscedastic error series (Hyndman and Athanasopoulos, 2018). Inputs were z-normalized by subtracting the mean and dividing by the standard deviation prior to being fed to the model to help stabilize the calculation of gradients during training. Models were implemented through custom scripts in R, benefiting from the “forecast” package (Hyndman and Khandakar, 2008; Hyndman et al., 2022). All models were utilized for multi-step ahead forecasting (e.g., 12-steps ahead forecasting at the monthly scale; except for the annual scale where only 1-step ahead forecasts were needed). A rolling window (with a step size equal to one) was used for concatenating the forecasted value at each time step and the previous inputs (lags and previous forecasted values) for ANN and SARIMA.

2.4.1.3 SARIMA

SARIMA is selected as the linear forecasting method in this work. The general form of a SARIMA(p, d, q) $\times$ (P, D, Q)_s model given the observation data (precipitation, y_t) is given as (Brockwell and Davis, 2016):

$$\phi(B)\Phi(B^s)(1-B^s)^D(1-B)^d y_t = \theta(B)\Theta(B^s)w_t, w_t \sim N(0, \sigma^2) \quad \text{Eq 2-2}$$

where s is the period, B is the backshift operator ($B^K y_t = y_{t-K}$), $\phi(B) = 1 - \phi_1 B - \dots - \phi_p B^p$, $\Phi(B^s) = 1 - \Phi_1 B^s - \dots - \Phi_P B^{sP}$, $\theta(B) = 1 + \theta_1 B + \dots + \theta_q B^q$, $\Theta(B^s) = 1 + \Theta_1 B + \dots + \Theta_Q B^{sQ}$. The values D and d are the order of seasonal and first differentiation, respectively. The values of D and d are obtained through the approach suggested by Hyndman and Athanasopoulos (2018) using the strength of seasonality (Wang et al., 2006) computed from seasonal and trend decomposition using loess regression (STL decomposition, Cleveland et al., 1990) and the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) unit root test (Kwiatkowski et al., 1992), respectively. The values p , q , P , and Q are tuned for each dataset (and for all aggregation levels) based on the AIC (Bozdogan, 1987) of the training set (1950-2004). The SARIMA model parameters are calculated by minimizing the conditional sum of squares (Brockwell and Davis, 2016). Model selection, model fitting, and forecasting were carried out using the “forecast” package (Hyndman and Khandakar, 2008; Hyndman et al., 2022) in the R environment.

2.4.2 Temporal Hierarchical Reconciliation Framework

Considering $y_{t=1,\dots,N}$, where N is the number of observations at the highest sampling frequency per year, m (which is set to 12), k aggregates can be constructed, where k is a factor of m . All q factors of m are denoted in descending order as $k \in \{k_q = m, \dots, k_1 = 1\}$. As an example, for monthly data (highest frequency), $k \in \{12, 6, 4, 3, 2, 1\}$. Each aggregation level ($\mathbf{y}_i^k$) is defined by a vector (Athanasopoulos et al., 2017):

$$\mathbf{y}_i^k = (y_{M_k(i-1)+1}^k, y_{M_k(i-1)+2}^k, \dots, y_{M_k i}^k)' \quad \text{Eq 2-3}$$

where the index $i = 1, 2, \dots, \lfloor T/m \rfloor$, and $M_k = m/k$. By stacking all the vectors in one column, results in (Athanasopoulos et al., 2017):

$$\mathbf{y}_i = (y_i^{k_q=m}, y_i^{k_{q-1}}, \dots, y_i^1)' \quad \text{Eq 2-4}$$

Eq 2-4 can be written as (Athanasopoulos et al., 2017):

$$\mathbf{y}_i = \mathbf{S} \mathbf{y}_i^1 \quad \text{Eq 2-5}$$

where $\mathbf{S}$ is the summation matrix. In general, $\mathbf{S}$ is a $(K = \sum_{j=1,\dots,q} k_j) \times m$ matrix. As an example, $\mathbf{S}$ for $m=12$ is given by (Nystrup et al., 2020):

$$\mathbf{S} = \begin{bmatrix} \mathbf{I}_1 \otimes \mathbf{1}_{12} \\ \vdots \\ \mathbf{I}_{12} \otimes \mathbf{1}_1 \end{bmatrix} \quad \text{Eq 2-6}$$

where $\mathbf{I}_i$ is the identity of order i , $\mathbf{1}_j$ is a length- j vector of ones, and $\otimes$ is the Kronecker product (Werner et al., 2008). The hierarchical representation introduced in Eq 2-6 is used to produce more accurate forecasts than those generated by the BU method. The hierarchical structure is demonstrated in Figure 2-2. To do so, the constraint that any forecast at an aggregate level is equal to the sum of the forecasts of the respective sub-aggregate forecasts is imposed (Athanasopoulos et al., 2017).

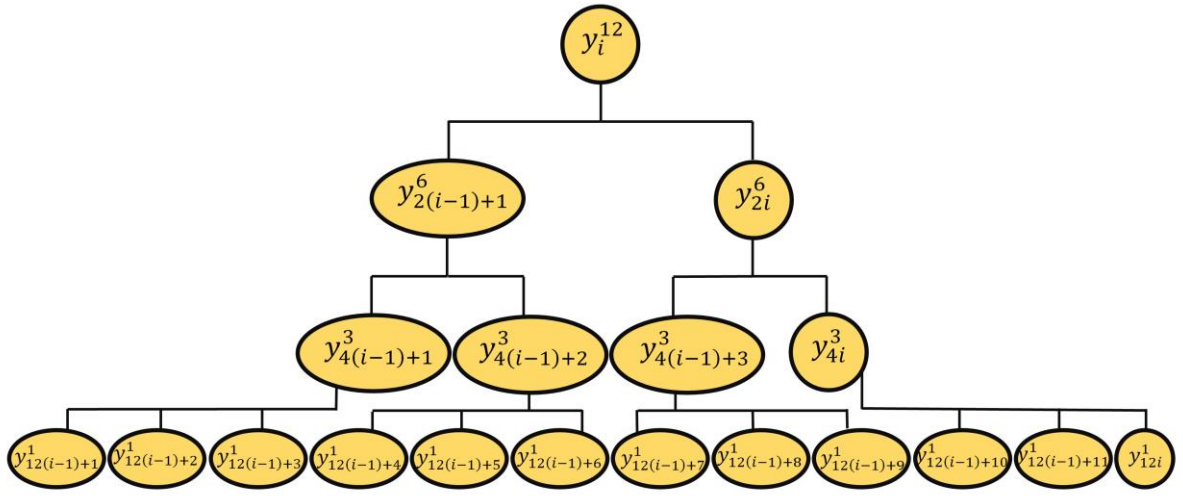


Figure 2-2. Temporal hierarchy for monthly data using the same index (i) for all aggregation levels. $i = 1, 2, \dots, [T/12]$, where T , is the number of observations. It should be noted that the structure below is only one of the possible presentations for aggregating monthly data with a yearly forecast horizon. This figure is adopted from Athanasopoulos et al. (2017).

Assuming h^* to be the maximum required forecast horizon at the most disaggregated level (monthly), $h = 1, \dots, [h^*/m]$ states the forecast horizon at the most aggregated level (annual). For clarity, these forecasts are referred to as base forecasts or original forecasts (as mentioned in Section 2.2). Eq 2-3-Eq 2-5 can be adopted for stacking the original h -step ahead forecasts (Athanasopoulos et al., 2017):

$$\hat{\mathbf{y}}_i(h) = (\hat{\mathbf{y}}_i^{k_q=m}(h)', \hat{\mathbf{y}}_i^{k_q-1}(h)', \dots, \hat{\mathbf{y}}_i^1(h)')' \quad \text{Eq 2-7}$$

where $\hat{\mathbf{y}}_i^k(h)$ is the h -step ahead forecast vector at the k^{th} level at time i , respectively. The regression representation of forecasts based on THR is given as (Athanasopoulos et al., 2017):

$$\hat{\mathbf{y}}_i(h) = \mathbf{S}\boldsymbol{\beta}_i(h) + \varepsilon_h \quad \text{Eq 2-8}$$

where ε_h is the THR error with 0 mean and unknown covariance matrix $\boldsymbol{\Sigma}_h$, and $\boldsymbol{\beta}_i(h)$ is the unknown conditional mean of the future values at the most disaggregated level (Athanasopoulos et al., 2017):

$$\boldsymbol{\beta}_i(h) = \mathbb{E}[\mathbf{y}_{\lfloor \frac{T}{m} \rfloor + h}^1 | y_1, y_2, \dots, y_T] \quad \text{Eq 2-9}$$

If $\boldsymbol{\Sigma}_h$ is known, generalized least squares can be used to derive $\boldsymbol{\beta}_i(h)$ (Athanasopoulos et al., 2017):

$$\boldsymbol{\beta}_i^{\text{GLS}}(h) = (\mathbf{S}'\boldsymbol{\Sigma}_h^{-1}\mathbf{S})^{-1}\mathbf{S}'\boldsymbol{\Sigma}_h^{-1}\hat{\mathbf{y}}_i(h) \quad \text{Eq 2-10}$$

using Eq 2-10, the reconciled forecasts are given by (Athanasopoulos et al., 2017):

$$\tilde{\mathbf{y}}_i(h) = \mathbf{S}\boldsymbol{\beta}_i^{\text{GLS}}(h) \quad \text{Eq 2-11}$$

2.4.2.1 Temporal Hierarchical Reconciliation Regression Estimation

Wickramasuriya et al. (2019) proved that $\boldsymbol{\Sigma}_h$ is not identifiable, and an alternative estimation should be used. Three different methods are applied for estimating $\boldsymbol{\Sigma}_h$ in this study. In the first method, the ordinary least squares, or OLS, method, $\boldsymbol{\Sigma}_h$ is assumed to be proportional to the identity matrix ($\mathbf{I}$) (Hyndman et al., 2011), therefore:

$$\boldsymbol{\beta}_i^{\text{OLS}}(h) = (\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'\hat{\mathbf{y}}_i(h) \quad \text{Eq 2-12}$$

The second approach, which is the series variance scaling-based weighted least squares, or WLS_v, method, considers a homogeneous error variance for each timescale (Athanasopoulos et al., 2017):

$$\boldsymbol{\Sigma}_h \approx \boldsymbol{\Lambda}_v = \text{diag}(\hat{\sigma}^{k_q}, \underbrace{\hat{\sigma}^{k_q-1}, \dots, \hat{\sigma}^{k_q-1}}_{M_{k_q-1}}, \dots, \underbrace{\hat{\sigma}^1, \dots, \hat{\sigma}^1}_m)^2 \quad \text{Eq 2-13}$$

where $\hat{\sigma}^k$ is the in sample 1-step ahead forecast error standard deviation at the k^{th} aggregation level.

By inserting Eq 2-13 in Eq 2-10 (Athanasopoulos et al., 2017):

$$\boldsymbol{\beta}_i^{\text{WLS}_v}(h) = (\mathbf{S}'\boldsymbol{\Lambda}_v^{-1}\mathbf{S})^{-1}\mathbf{S}'\boldsymbol{\Lambda}_v^{-1}\hat{\mathbf{y}}_i(h) \quad \text{Eq 2-14}$$

The third method estimates Σ_h purely based on the hierarchical structure (Athanasopoulos et al., 2017), and is named structural scaling-based weighted least squares (WLS_S). In WLS_S instead of utilizing the in-sample error, it makes use of the hierarchical structure of the data (Athanasopoulos et al., 2017):

$$\Sigma_h \approx \Lambda_s = \text{diag}(k_q = m, \underbrace{k_q - 1, \dots, k_q - 1}_{M_{k_q-1}}, \underbrace{k_1 = 1, \dots, 1}_m) \quad \text{Eq 2-15}$$

similar to Eq 2-14, the following expression can be given (Athanasopoulos et al., 2017):

$$\beta_i^{\text{WLS}_S}(h) = (\mathbf{S}'\Lambda_s^{-1}\mathbf{S})^{-1}\mathbf{S}'\Lambda_s^{-1}\hat{\mathbf{y}}_i(h) \quad \text{Eq 2-16}$$

The three introduced estimation methods are employed to obtain the temporal hierarchical reconciled forecast at each timescale. Alongside the THR methods, BU method is also applied to the forecasts. The OLS, WLS_V, WLS_S, and BU models' accuracy are compared to each other and to the ORFs. Most of the calculations related to THR are done through the “thief” package (Athanasopoulos et al., 2017; Hyndman and Kourentzes, 2018).

2.4.3 Performance Measures

Several performance measures were adopted to evaluate the different forecasts. Nash-Sutcliffe efficiency (NSE; Nash and Sutcliffe, 1970) is used as a scale-independent performance measure, while the normalized mean absolute error (nMAE) and normalized root mean square error (nRMSE) indices are adopted to allow the generalization of forecasting performance across the basins. Normalization is applied by dividing MAE and RMSE by the means of the targets (observations). For identifying whether the THR framework has significantly improved forecast performance compared to the benchmark (ORFs) or not, the Diebold-Mariano (DM) test is adopted (Diebold and Mariano, 2002; Harvey et al., 1997). The DM test compares the accuracy of two sets of forecasts based on a function of the residuals. Usually, accuracy is measured based on the square of the errors (for more details, please see Harvey et al., 1997), which is referred to as the DM-2 test. As the main reason to employ a THR framework is to obtain consistent forecasts at different aggregation levels, minimizing the forecasts' total deviation from the observations at the end of the forecast horizon is also desirable. Therefore, besides the square of errors, absolute errors are also used in the DM test. The latter is referred to as the DM-1 test.

2.5 Result and Discussion

This section provides the base forecasts' (ORFs') results, which serve as the benchmark for the reconciled forecasts. All the presented results are associated with the test set (2005–2010). It should be noted that for all timescales, the three adopted methods (ETS, ANN, SARIMA) were trained/calibrated for 1-step ahead forecasting, while in the testing phase, using a rolling window, models were applied for multi-step ahead forecasting (e.g., 12-steps ahead forecast for the monthly scale). For ETS, direct multi-step ahead forecasting was implemented, while for ANN and SARIMA, a recursive input window was used for forecasting (please see Ben Taieb et al., 2012). On average (for each basin and for all timescales), optimization/training of ETS models took 3.28 s (standard deviation = 0.69 s), ANN models took 16.14 s (standard deviation = 3.34 s), and SARIMA models took 19.33 s (standard deviation = 4.77 s). The obtained times were recorded on a workstation with an Intel(R) Core (TM) i7-6700HQ CPU @ 2.60 GHz and 12 GB RAM. Error assessment associated with the training set (1-step ahead forecasting for all models) is provided in Appendix A.

2.5.1 ORF Accuracy

At each of the six designated timescales, forecasts were produced with an annual forecast horizon (resulting in a 12-steps (monthly), 6-steps (2-monthly), 4-steps (quarterly), 3-steps (4-monthly), 2-steps (bi-annual), and 1-step (annual) ahead forecast) for all 84 selected basins using ETS, ANN, and SARIMA models. The residual plots of monthly forecasts (12-steps ahead) for each of the forecast models are depicted in Figure 2-3. The results indicate that the forecast models' performances were not similar across the different basins. However, Figure 2-3 shows that the average error (across all basins) of each method is close to 0 (average nRMSE (–) values are 0.484, 0.523, and 0.526 for ETS, ANN, and SARIMA, respectively). Figure 2-4 depicts the fitted kernel density estimate (using a Gaussian kernel with a bandwidth equal to 0.1, see Sheather (2004) for details) of the probability density function (pdf) of NSE values obtained by ETS at each timescale to better summarize the variation in forecast accuracy across the various basins. The results confirm that not only are the forecast models' performance dissimilar across basins but also across different timescales. For the ETS model, it is revealed that a large proportion of forecasts at bi-annual and annual timescales have resulted in NSE values smaller than 0, which are considered unsatisfactory and poorer forecasts than using the mean of the observed precipitation record as the forecast (Moriasi et al., 2007).

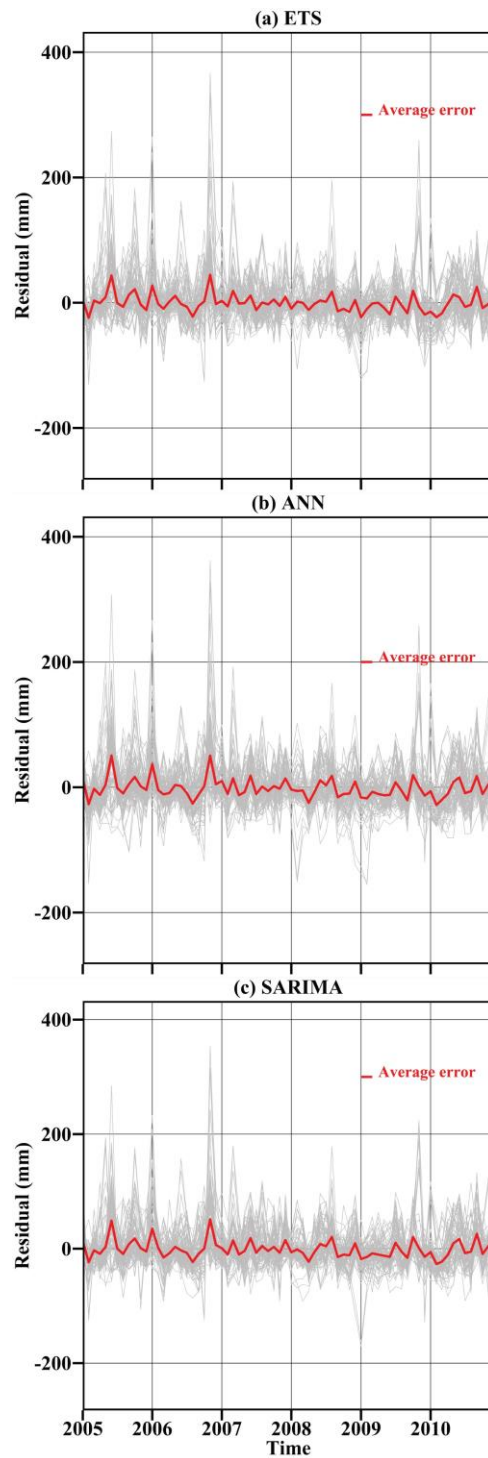


Figure 2-3. Residual plot of original forecasts for the three different models: (a) exponential time-series smoothing (b) artificial neural network, and (c) seasonal auto-regressive integrated moving average at the

monthly scale (i.e., 12-steps ahead forecasts) from 2005 to the end of 2010 for 84 selected basins. Each gray line represents the residuals (of the forecasts) for a specific station. The average error across all basins is shown with a bold solid line.

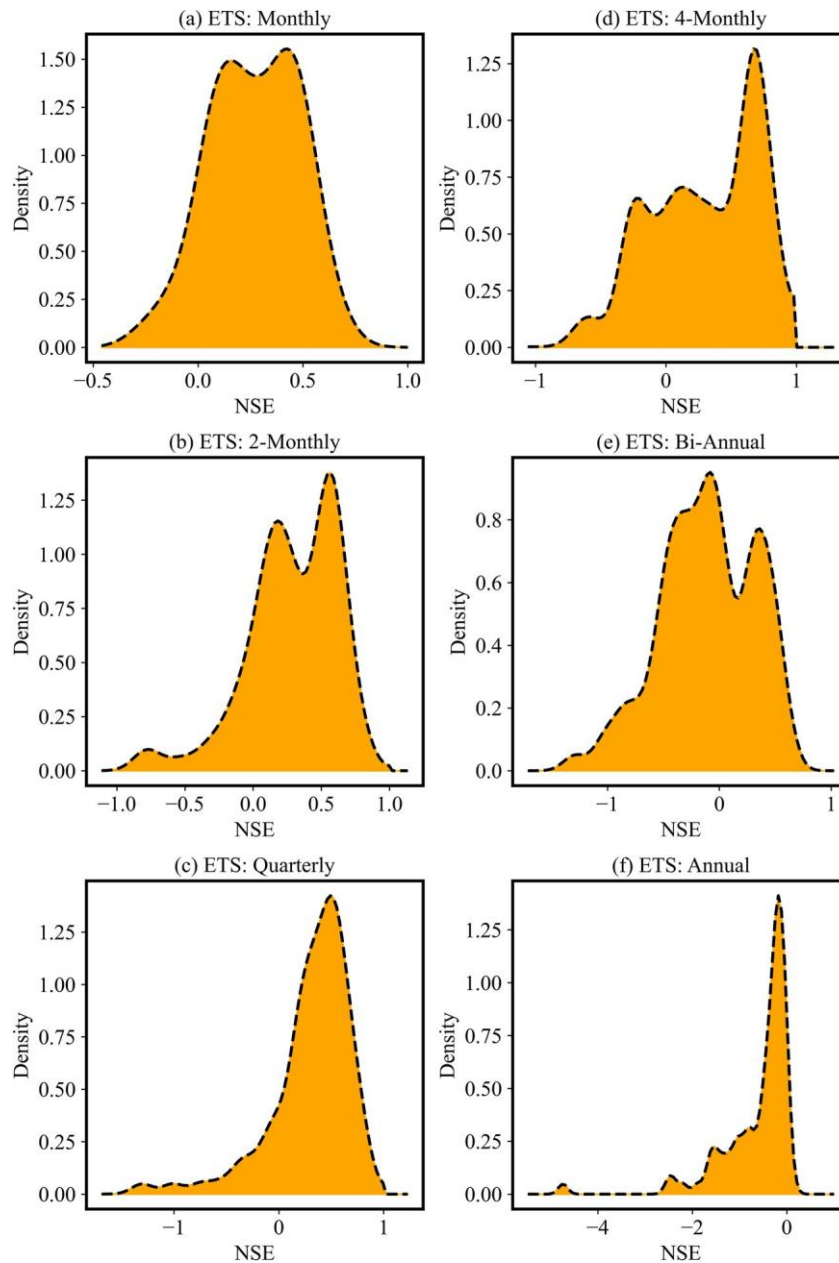


Figure 2-4. The non-parametric fitted probability density function (pdf) of the Nash-Sutcliffe efficiency values associated with the exponential time-series smoothing model at each timescale. The area under the dashed line (kernel density estimate) represents the area under the pdf curve (probability).

The number of obtained forecasts at each timescale with NSE (–) greater than 0, average NSE across basins with NSE greater than zero, and average NSE scores across all basins are listed in Table 2-2. The results show that for all timescales except the annual scale, which included only a single basin where the NSE was greater than 0, ETS has the highest number of forecasts with NSE greater than 0 and the highest average NSE (across all timescales). For the annual scale, the highest average NSE results from the number of samples at this scale (one remaining forecast, see Table 2-2). It can also be seen that, in general, the SARIMA method outperforms ANN. However, although ANN produced fewer forecasts with NSE greater than 0 at the monthly scale when compared to SARIMA, the basin-averaged NSE of those forecasts was higher (for ANN). Based on the basin-averaged NSE of the filtered results (NSE greater than zero), for all three forecast models, accuracy increases as the scale of temporal aggregation is increased from monthly up to 4-monthly (with 4-monthly having the highest accuracy) and then decreasing for bi-annual and annual scales. This observed pattern could be related to the highly seasonal nature of precipitation in the study areas. Aggregating the time-series up to the 4-month timescale has resulted in a “deseasonalized” time-series, leading to a higher level of “predictability.” Temporal aggregation acts as a filter, attenuating seasonal and random components, and therefore, resulting in smoother time-series (Kourentzes and Petropoulos, 2016). The relatively small data sample size at the annual scale is likely the main contributor to the forecast models’ inferior performance for this timescale. Although ANN models (as nonlinear models) are typically considered appropriate for water resources forecasting (Papacharalampous et al., 2019), ETS and SARIMA both outperformed the ANN method here. The main reason for this is that ANN requires more samples to estimate the nonlinear relationship between past observations and future values. Also, careful hyperparameter adjustment (the number of neurons/nodes in the hidden layer, the number of training epochs, learning rate, type of activation function, etc.), which is not thoroughly explored in automatic forecasting routines (such as the one adopted here) is important for ANN models. The experiments in this study showed that careful consideration should be taken when setting the training termination policy (absolute or relative change in consecutive training epochs) for ANN models. Finally, the ORF results also indicate that ETS models took considerably less time for optimization and their accuracy was higher than ANN and SARIMA.

Table 2-2. ORF analysis. Number of remaining datasets (from initial 84), the average NSE (-) value for each forecasting method after removing basins with NSE values (at the respective timescale) less than 0, and the average of NSE values across all basins

Forecast method	Timescale					
	Monthly	2-Monthly	Quarterly	4-Monthly	Bi-annual	Annual
ETS	77,0.305,0.269	71,0.394,0.291	70,0.440,0.304	63,0.480,0.295	31,0.325,-0.110	1,0.110,-0.614
ANN	62,0.246,0.158	59,0.376,0.219	64,0.380,0.211	55,0.487,0.222	26,0.232,-0.224	5,0.072,-0.72
SARIMA	66,0.212,0.154	66,0.372,0.268	69,0.418,0.289	56,0.526,0.265	27,0.295,-0.141	5,0.113,-0.745

2.5.2 Temporal Hierarchical Reconciliation

2.5.2.1 Impact on ORFs' Accuracy

Three THR methods, including WLS_s, WLS_v, and OLS, were applied for THR and compared against the BU method. For illustration purposes, the interested reader is directed to an example in Table A 2 in Appendix A, demonstrating that the proposed THR methods result in consistent forecasts.

DM-1 and DM-2 tests were employed for assessing any improvement in forecasting performance due to the THR methods (WLS_s, WLS_v, OLS) and the BU method when compared against the ORFs. The focus here is on the number of models that improve upon the ORFs' performance instead of the magnitude of change in model performance. The net percentage of improvement based on the number of WLS_s, WLS_v, OLS, and BU forecasts showing better performance than ORF according to DM-1 and DM-2 tests are depicted in Figure 2-5. The results confirm that THR has improved upon the ORFs at different timescales, but the pattern in these improvements is nuanced. For the ETS forecasts, it is observed that the annual timescale benefited the most from the THR, while for monthly and bi-annual timescales, THR lowered the accuracy. It is observed that, in general, the BU method outperformed the hierarchical methods when ETS was utilized for producing the ORFs. For both ANN and SARIMA, a similar pattern is observed. The BU method results in the poorest (forecast) reconciliation, while for hierarchical methods, the monthly series shows the most improvement through THR. It is observed that the net improvement decreases as the scale of aggregation increases (from monthly to annual), indicating that, for this case study, the lower levels of the temporal hierarchy benefited most from THR compared to the upper levels. Although this study's methodology is different from Rostami-Tabar et

al. (2013), the results presented here confirm that accuracy improvements can be gained by forecasting time-series at higher levels of aggregation and using those forecasts to improve the forecasts' accuracy at lower levels of aggregation.

In general, based on the DM-1 and DM-2 results, for ANN, the OLS method showed better performance compared to WLS_S and WLS_V, while for SARIMA, WLS_V showed better performance. WLS_S uses only the structure (relative position) of the time-series in the temporal hierarchy for reconciliation (please see Figure 2-2). It is also observed that OLS, in general, showed better performance than WLS_S and WLS_V for higher aggregation levels, such as 4-monthly, bi-annual, and annual. The results signify that less complex methods such as WLS_S and OLS could be used for THR without requiring in-sample error variance (as the error is not always available in practical cases), which corresponds to the findings of Athanasopoulos et al. (2017). However, these results contradict the findings of Yang et al. (2017), where they found WLS_V outperformed WLS_S, OLS, and BU when applied for forecasting daily solar power, which is related to the fact that in this study the various models did not perform well in the training phase (high in-sample error variances) and due to the difference in the time-frequency representation of daily solar power versus monthly precipitation.

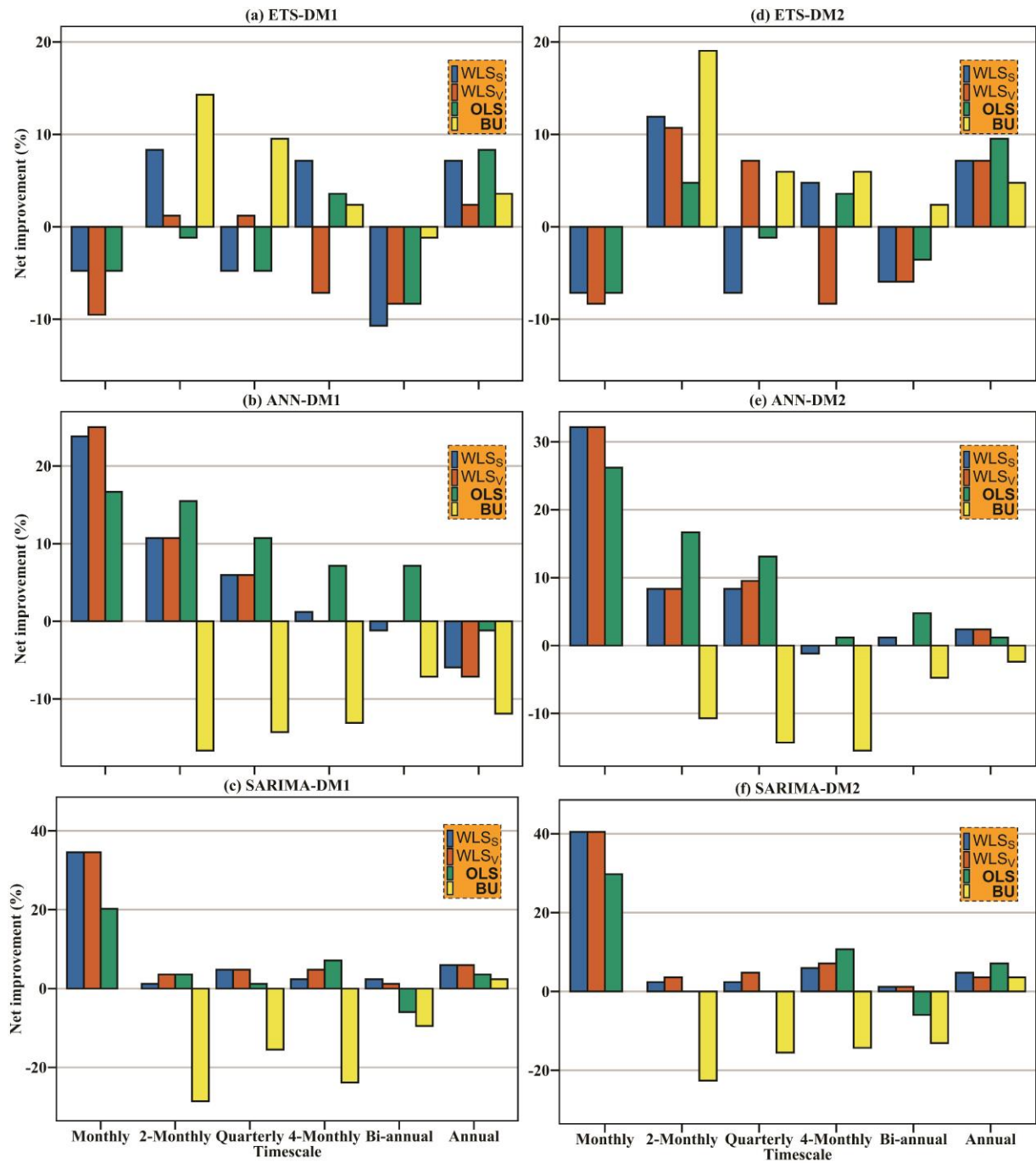


Figure 2-5. The net percentage of improvement (based on the number of forecasts) due to different temporal hierarchical reconciliation methods (WLS_S, WLS_V, OLS, and BU) with respect to the original forecasts (ORF) based on DM-1 and DM-2 tests. Note: BU is not reported for the monthly scale as it is equal to the ORF.

The previous section (Section 2.5.1) showed that the ORFs had dissimilar accuracy at different timescales. It is noted that inaccurate ORFs may have impacted the THR results. For example, as depicted in Figure 2-6, two ORFs were chosen based on their nRMSE value at the 4-monthly scale (since all forecast methods showed the highest accuracy at this timescale) along with the ORF values, the observed data (OBS), and the WLS_s and BU forecasts. The selected forecasts are for different basins, but each had an NSE value greater than 0. The best ORF had the lowest nRMSE while the mediocre ORF was selected based on the median rank among all the ORFs (e.g., rank 28 for SARIMA with 56 forecasts having NSE greater than 0, see also Table 2-2. ORF analysis. Number of remaining datasets (from initial 84), the average NSE (-) value for each forecasting method after removing basins with NSE values (at the respective timescale) less than 0, and the average of NSE values across all basins). The results indicated that accuracy for a selected basin is highly model-dependent. It is seen that temporal hierarchically reconciled (or simply reconciled) forecasts from different methods are very close to each other for the best performing models, especially ETS. However, the results become more dissimilar (except for ETS) as the forecasting performance degrades. The results confirm that the ORFs' accuracy has important impacts on the reconciled forecasts' accuracy. This finding was expected, especially for the WLS_v method, where the covariance error matrix is constructed based on the forecasts' in-sample error (see Section 2.4.2.1).

As noted earlier, forecast accuracy varies across timescales, this finding is discussed in more detail below. Besides obtaining consistent forecasts at different timescales, one primary purpose of THR is to leverage the more accurate forecasts to improve forecasts at less accurate timescales. The results demonstrate that for ANN and SARIMA, THR methods outperform the BU method. For ETS, however, it is observed that simply aggregating the monthly forecasts (i.e., via the BU method) results in more accurate forecasts. This finding is due to two reasons. First, the ETS method produces more accurate forecasts at the monthly scale than ANN and SARIMA as ETS's multi-step ahead forecasts were obtained directly at each timescale, rather than recursively. This results in better BU forecasts compared to ORF. Second, as the level of temporal aggregation increases, the performance of ANN and SARIMA significantly improves (especially when compared to ETS). The results indicate that when the BU method is more accurate than ORF (such as for the ETS method, in this case), the reconciliation accuracy may decrease, although THR methods result in consistent forecasts. As the THR methods use linear regression for reconciliation, a significant increase in the accuracy at higher levels of the temporal hierarchy will also improve the forecast accuracy at lower levels. It is noted that the impact of different

timescales on each other (in terms of forecast performance) is dependent on the error covariance matrix and the hierarchy structure (see Section 2.4 and Appendix A for additional discussion).

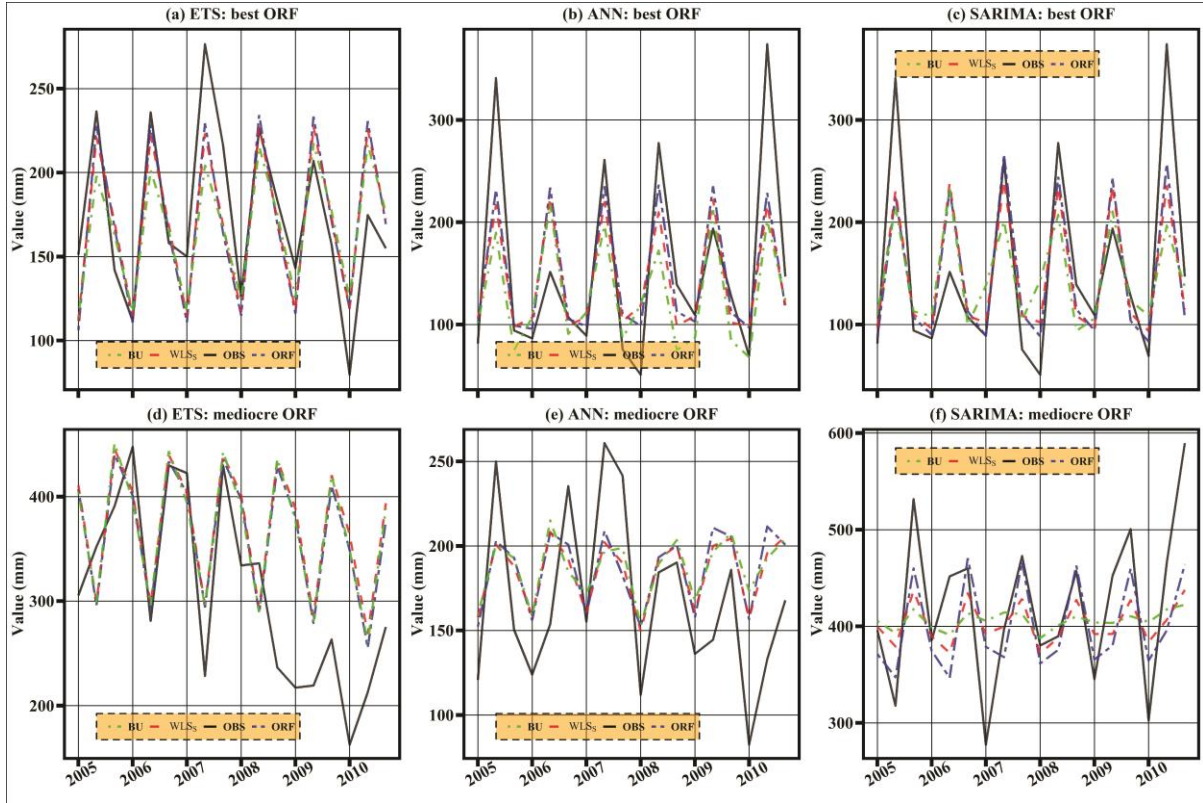


Figure 2-6. The original forecasts (ORFs), observed data (OBS), WLSs, and BU forecasts for two cases associated with the 4-monthly scale: first row figures (a, b, c) are associated with the best ORF based on the nRMSE (Canadian model parameter experiment [CANOPEX] IDs: 515, 25, and 25, respectively), while the second-row figures (d, e, f) are associated with “mediocre” forecasts (CANOPEX IDs: 422, 105, and 328, respectively).

2.5.2.2 Assessment of ORF Accuracy on Temporal Hierarchical Reconciliation

The THR results in Figure 2-5 were obtained using all forecasts, where the DM-1 and DM-2 results show how many forecasts were improved/worsened (but not by how much the forecast accuracy changed). To explore how the ORFs’ accuracy affects the reconciled results, the reconciled forecasts associated with ORFs that had NSE smaller than 0 (see Table 2-2) were omitted from the set of forecasts at each timescale (since these forecasts would not be used in practice as they perform worse than using the mean of the observed record as a forecast). Hereafter, these forecasts are referred to as filtered

forecasts. It is important to note that all timescales were used for THR; however, in this subsection, the analysis only considers the reconciled forecasts whose corresponding ORF have a NSE greater than zero. The percent (%) relative change in nRMSE values of the reconciled forecasts with respect to the ORFs for monthly, 2-monthly, and quarterly timescales, are illustrated in Figure 2-7 using violin plots, where the circle marker and the solid line denote the mean and the two standard deviation range, respectively. In Figure 2-7, values below 0 indicate a relative improvement in the nRMSE. The results indicate that the reconciled forecasts through the BU method have the highest spread. It is also observed that some extreme cases heavily influence the average values; for example, considering SARIMA at the quarterly scale, the BU method increased the relative error up to approximately 40% (Figure 2-7i). In general, for ETS forecasts, THR did not significantly impact the model performance (Figure 2-7 (a), (d), (g)), where the BU method slightly outperformed the THR methods. For ANN and SARIMA forecasts, THR methods (WLS_s, WLS_v, and OLS) outperformed the BU method, especially at the monthly timescale (where BU is equal to ORF). As the time-series becomes more coarsely aggregated, it is observed that the relative nRMSE values' distribution for these methods becomes symmetrical (around the two standard deviation range) with a mean close to 0, which shows that it cannot be known a priori whether THR will improve forecast performance at a given basin, highlighting the importance of carrying out large-sample experiments when testing and validating new forecasting approaches (Papacharalampous et al., 2020), such as THR. The results are aligned with the Global Energy Forecasting Competition 2017 (GEFCom2017), where using hierarchical information did not result in a substantial improvement compared to utilizing hybrid forecast models (Hong et al., 2019).

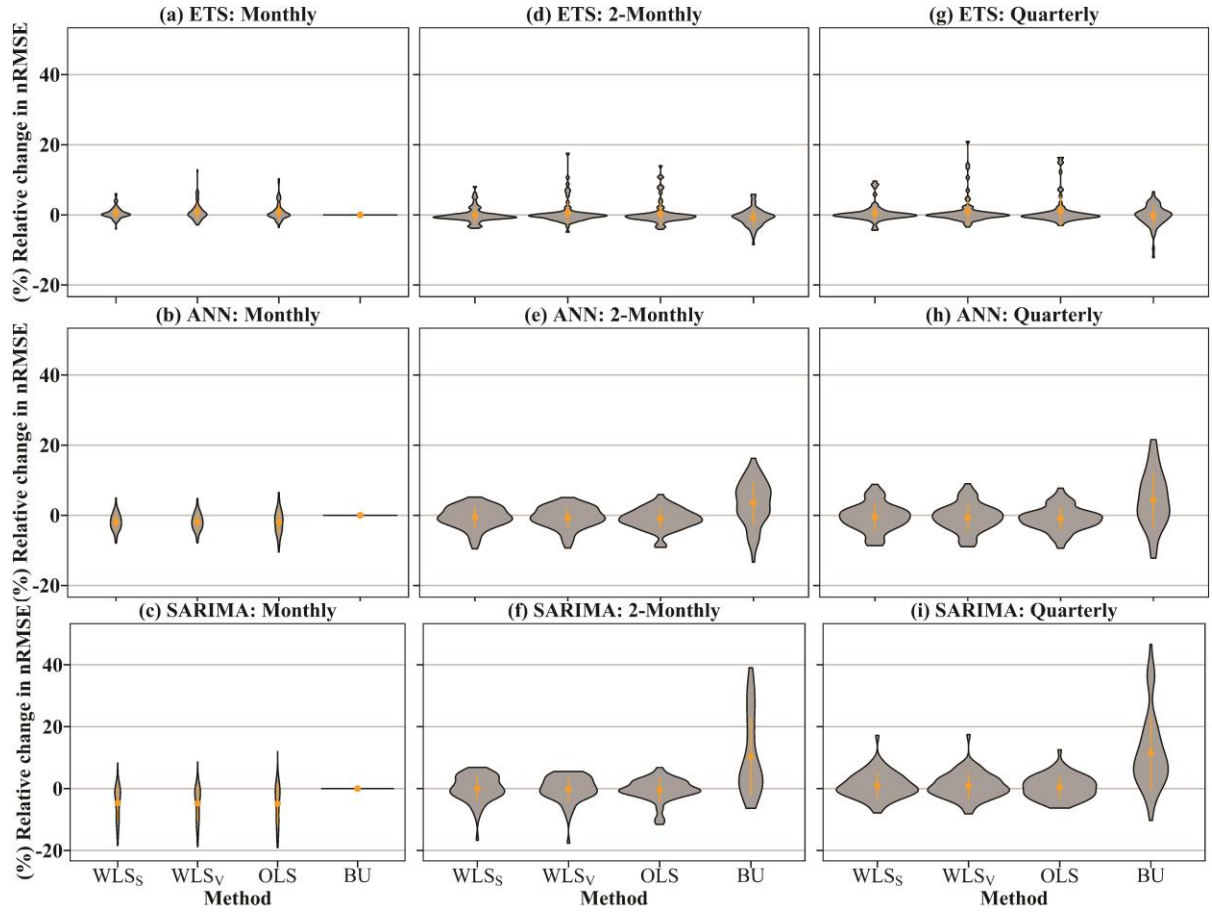


Figure 2-7. Violin plots representing the dispersion of relative $([\text{reconciled-ORF}]/\text{ORF} \times 100\%)$ nRMSE values of the temporal hierarchical reconciliation methods compared to original forecasts (ORFs) for monthly (first column), 2-monthly (second column), and 4-monthly (third column) timescales. Bottom-Up (BU) is equal to ORF for the monthly scale. The circle marker and the solid line show the mean and the two standard deviation range ($\text{mean} \pm 2 \times \text{standard deviation}$), respectively. The results are associated with filtered ORFs.

The average normalized mean absolute error (nMAE) (–) values obtained through the THR methods compared to ORFs are listed in Table 2-3. It is noted that the results are associated with the filtered forecasts. The obtained nMAE values agree with the nRMSE values depicted in Figure 2-7. It is shown that the effect of THR was not as substantial for ETS as it was for ANN and SARIMA forecasts (especially for the monthly timescale). Further, the results confirm that the BU method substantially reduces ORF accuracy when coupled with ANN and SARIMA. The results show that the THR methods can be used as an appropriate alternative to the BU method, resulting in consistent forecasts and

improving the forecasts' accuracy. Like Table 2-3, the Appendix A includes a table reporting NSE values (instead of nMAE).

Table 2-3. Average nMAE (-) values for ORF and the temporal hierarchical reconciliation methods obtained by each model (ETS, ANN, and SARIMA). Based on the definition of BU, the monthly scale values listed for ORF and BU are equal. At each timescale, the results are associated with ORFs having an NSE greater than 0.

	Monthly	2-Monthly	Quarterly	4-Monthly	Bi-annual	Annual
Method	ETS					
ORF	0.366	0.275	0.235	0.206	0.162	0.131
WLS _S	0.367	0.275	0.236	0.206	0.163	0.133
WLS _V	0.368	0.277	0.237	0.207	0.162	0.132
OLS	0.368	0.277	0.238	0.207	0.162	0.132
BU	0.366	0.274	0.235	0.207	0.164	0.131
Method	ANN					
ORF	0.401	0.299	0.255	0.221	0.172	0.102
WLS _S	0.392	0.297	0.254	0.226	0.168	0.102
WLS _V	0.392	0.297	0.254	0.226	0.168	0.102
OLS	0.391	0.295	0.252	0.222	0.168	0.102
BU	0.401	0.312	0.268	0.244	0.175	0.126
Method	SARIMA					
ORF	0.403	0.288	0.242	0.204	0.155	0.121
WLS _S	0.384	0.288	0.244	0.213	0.157	0.127
WLS _V	0.383	0.287	0.244	0.212	0.157	0.127
OLS	0.382	0.285	0.242	0.209	0.157	0.125
BU	0.403	0.319	0.274	0.253	0.168	0.124

It is important to remember that although consistent forecasts at different timescales may be useful for various purposes (e.g., long-term planning, short-term operations), the forecast accuracy at some timescales may be more valuable than others depending on the intended use of such forecasts. For example, suppose the main objective is to produce accurate monthly forecasts (12 step-ahead forecasts), but it is also important that forecasts be consistent at all other five timescales (2-monthly up to annual). In that case, it is possible to assess overall improvements due to THR by filtering forecasts according to the ORFs' accuracy (NSE) at the monthly scale. In other words, only the ORFs that have an NSE less than 0 at the monthly scale would be removed from further consideration (even if their forecasts'

NSE at higher timescales are less than 0, since the monthly scale is given precedence in the evaluation of the different models). The main difference between the latter evaluation and the results provided in Figure 2-7 is that by considering the monthly scale as the main forecast horizon of interest, 77, 62, and 66 basins would be considered for reconciliation for all timescales using ETS, ANN, and SARIMA, respectively. In the analysis that follows, the nRMSE values of the ORFs and their reconciled versions (WLSS, WLSV, OLS, and BU) for all basins (having ORFs with NSE greater than 0 at the monthly-scale) were sorted in increasing order based on the ORFs' nRMSE; the basin-averaged results for each group of forecasts (the combination of each forecasting model and reconciliation method) are illustrated in Figure 2-8 (see also Fig A 3 in Appendix A). The x-axis shows the accumulated percentage of accurate monthly forecasts (i.e., 20% is associated with the top 20 percentile monthly ORFs, while 100% corresponds to all ORFs, with NSE greater than 0 at the monthly scale). Values for the % nRMSE change lower than 0 indicate an overall accuracy improvement due to the reconciliation method. Relative changes (%) in nRMSE values based on ORFs with NSE greater than 0 at each timescale are depicted in Figure 2-7. Figure 2-8, however, illustrates the relationship between the ORF accuracy at the monthly scale (the most disaggregated series, and not individually filtered for all timescales as in Figure 2-7) and the performance of each reconciliation method (WLS_s, WLS_v, OLS, and BU).

In general, for ANN and SARIMA, it is observed that the THR methods resulted in more accurate forecasts compared to BU. It can also be seen, for ANN and SARIMA, that less accurate ORFs at the monthly scale were more significantly improved due to reconciliation than those with relatively higher accuracy, where the top 20% ORFs showed almost no relative improvement in their nRMSE after reconciliation. In general, for other timescales, reconciled forecasts associated with more accurate forecasts at the monthly scale (e.g., top 20 percentile ORFs in Figure 2-8) improved more compared to others. The same pattern is observed for ETS, where except for the annual timescale, the performance of the reconciliation methods slightly decreases when the ORFs' accuracy decreases. Figure 2-8 and Fig A 3 in Appendix A corroborate the earlier described findings (see Section 2.5.2.1) that the BU method generally performs superior to the THR methods when combined with ETS, especially for less accurate forecasts at the monthly level. More details regarding the comparison between reconciliation methods are provided in Appendix A.

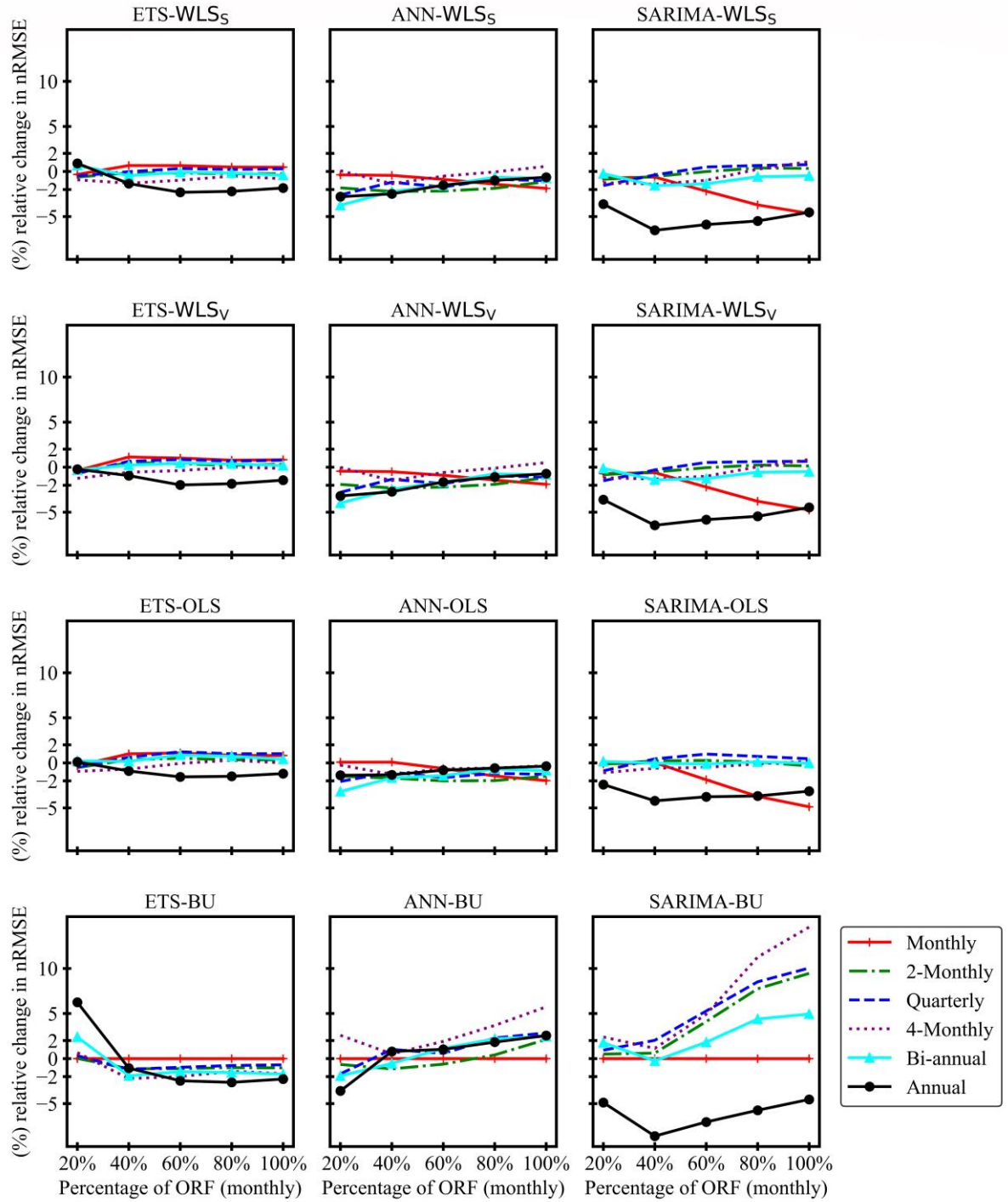


Figure 2-8. The relative normalized root mean square error (nRMSE) change of the temporal hierarchical reconciliation methods compared to original forecasts (ORFs) at different timescales ($[\text{reconciled-ORF}]/\text{ORF} \times$

100%). The results are associated with ORFs having a Nash-Sutcliffe efficiency value greater than zero at the monthly scale. The x -axis shows the accumulated percentage of sorted monthly forecasts (based on the nRMSE). Note: that bottom-up is equal to ORF at the monthly scale resulting in no relative change.

2.5.3 Temporal Hierarchy's Structure and Its Impact on Reconciliation

THR methods (WLS_S, WLS_V, OLS) as well as BU were implemented based on a hierarchy with six timescales (monthly, 2-monthly, quarterly, 4-monthly, bi-annually, and -annually). However, it may be of interest to only consider limited timescales (e.g., monthly, 2-monthly, and 4-monthly) with a custom forecast horizon (e.g., annual forecasting). In the new setup, forecasts were reconciled by considering only three timescales: monthly, 2-monthly, and 4-monthly (i.e., quarterly, bi-annual, and annual forecasts were not used for reconciliation). The forecast horizon in the new setup was kept as one year (i.e., 12-, 6-, and 3-steps ahead forecasts were considered at the monthly, 2-monthly, and 4-monthly scales, respectively). Average NSE values (filtered based on NSE values of ORFs) of the new reconciled set of forecasts are provided in Appendix A. For ETS forecasts, the NSE values were slightly increased for monthly and 2-monthly scales. However, for ANN and SARIMA, the results indicated that the accuracy of the THR methods slightly decreased by reducing the number of timescales involved in reconciliation. These results demonstrate that involving “inaccurate” forecasts (such as bi-annual and annual forecasts, in this case) in the reconciliation framework could improve forecast performance. Furthermore, the results show that the ORFs accuracy (as influenced by the forecast horizon) has a greater impact on reconciliation behavior compared to the temporal hierarchy structure used for reconciliation.

2.6 Summary, Conclusions, and Future Work

In this research, THR methods were investigated to understand how they could be used for obtaining consistent precipitation forecasts (see Section 2.1) and whether they could be utilized for improving the base forecasts' (ORFs') accuracy, using a large-scale experiment in Canada. Three THR methods, namely, WLS_S (structural scaling-based WLS), WLS_V (series variance scaling-based WLS), and OLS (Ordinary least squares), along with the simple BU (Bottom-Up) method, were considered for “reconciling” the forecasts. Three different forecasting methods, including ETS, ANN, and SARIMA, were selected to produce the ORFs. The forecast horizon was set as 1 year and forecasts were produced using a rolling window at monthly, 2-monthly, quarterly, 4-monthly, bi-annual, and annual scales. The different forecasting and THR methods were used to generate precipitation forecasts for selected basins

from the Canadian model parameter experiment (CANOPEX) dataset (see Section 2.3). The results highlighted that the forecast accuracy is basin-, timescale-, and forecast model-dependent.

It was highlighted that forecasting accuracy increases monotonically from the monthly to the 4-month scale, followed by a decreasing pattern from the 4-month to the annual scale. In general, ETS forecasts were more accurate than ANN and SARIMA. The results demonstrated that for ANN and SARIMA, the BU method had the worst performance among all four reconciliation methods considered in this study (especially for SARIMA forecasts, where the BU method displayed significantly poorer performance). Oppositely, for ETS, the BU method showed superior performance compared to the THR methods. The results indicated that both the WLS_S method (where the error covariance matrix is solely based on the temporal hierarchical structure the position of the time-series data relative to other aggregated/disaggregated series; see Section 2.4) and the OLS method slightly outperformed the WLS_V method. This finding is due to the difference between training error and test set error, where it was observed for basins in which the difference between training error and testing error was high (i.e., an indicator of model overfitting), the accuracy of the forecasts decreased when adopting the WLS_V method. Although, in general, it was observed that both accurate and inaccurate forecasts could be improved by THR, the results indicated that ORF accuracy had a much greater impact on the final reconciled forecasts compared to the adopted THR method. Therefore, different timescales (with different levels of accuracy) showed dissimilar sensitivity to THR.

In this work, the forecasts were generated using three different models commonly applied for time-series forecasting. However, all forecasts were issued for all six timescales using a single forecast model type (either ETS, ANN, or SARIMA). Future research should consider using the best possible forecast at each timescale using different forecast models and then apply THR to these “best” models. Furthermore, simultaneous spatial and temporal reconciliation (cross-temporal reconciliation; Fonzo and Girolimetto (2021)) as a post-processing technique could be leveraged for improving the accuracy of simulations across hydrological response units. Another potentially promising research avenue is coupling new stochastic water resources forecasting frameworks based on machine learning (Papacharalampous et al., 2019; Quilty and Adamowski, 2020) with probabilistic THR methods (Athanasopoulos et al., 2020; Jeon et al., 2019; Yagli et al., 2019) to obtain stochastic temporal hierarchically reconciled forecasts. Investigating the latter is valuable as uncertainty should be represented in water resources forecasting, and the benefits of utilizing hierarchical information compared to utilizing more complex probabilistic forecasting models is not yet clear (Hong et al., 2019).

Although several works have investigated THR, choosing the best method is still inconclusive and regarded as case- and timescale-dependent. The latter highlights a promising opportunity for a thorough investigation of THR for forecasting key water resources variables (e.g., streamflow, evaporation, urban water demand) whose consistent forecasts could be highly valuable for reliable decision-making.

Chapter 3

Generative Deep Learning for Probabilistic Streamflow Forecasting: Conditional Variational Auto-Encoder

This chapter is derived from two papers published in Journal of Hydrology. Both papers have been reorganized to align with the thesis format, resulting in some differences between this chapter and the original publications.

Jahangir, M.S. and Quilty, J., 2024. Generative deep learning for probabilistic streamflow forecasting: Conditional variational auto-encoder. *Journal of Hydrology*, 629, p.130498.

Jahangir, M.S., You, J. and Quilty, J., 2023. A quantile-based encoder-decoder framework for multi-step ahead runoff forecasting. *Journal of Hydrology*, 619, p.129269.

Summary

Probabilistic hydrological forecasting has gained increasing importance in recent years, as it offers essential information for risk-based decision-making and flood management. Traditional hydrological models often produce deterministic forecasts, which do not account for the inherent uncertainties in hydrological systems. Although many previous studies have investigated using deep learning (DL) models for hydrological prediction, the development of probabilistic DL models (especially, generative models) has not yet been thoroughly examined for hydrological forecasting. The present study investigates the efficacy of a generative DL model, namely, conditional variational auto-encoder (CVAE). CVAE is applied for one-seven day(s) ahead probabilistic streamflow forecasting in 75 basins from the Canadian model parameter experiment (CANOPEX) database. The CVAE forecast model, which outputs forecasts in the form of a probability distribution, was benchmarked against two state-of-the-art quantile-based DL models: quantile-based encoder-decoder (ED) and quantile-based CVAE (QCVAE). The latter outputs forecasts for specific quantiles of a probability distribution (here, $q = 0.05, 0.5, 0.95$). More than 9000 models were developed based on different basins, input variable sets, and model structures. The models were evaluated in terms of point forecast accuracy and forecast reliability. The results indicate that the CVAE model generally outperforms the benchmark models in terms of reliability at a 90 % confidence level (median reliability of 92.49 % compared to 87.35 % and 84.59 % for ED and QCVAE, respectively). However, the quantile-based forecast models produce slightly more accurate point forecasts than the CVAE (median Kling-Gupta efficiency (KGE) of 0.88

compared to 0.90 for both ED and QCVAE). Notably, the CVAE model exhibits superior probabilistic forecasts in basins with poor point forecast accuracy, highlighting its usefulness over benchmark methods in difficult-to-forecast basins. Overall, CVAE is a promising probabilistic DL model for streamflow forecasting, and it can be readily applied for forecasting other hydrological variables (evapotranspiration, water level, etc.). The findings of this study provide a basis for future research on probabilistic forecasting of hydrological variables using generative DL models.

3.1 Introduction

Accurate and reliable streamflow/runoff forecasting is critical for various water management tasks (Ghobadi and Kang, 2022), including water supply management, flood early warning systems, hydropower generation, and agricultural development. From an operational viewpoint, streamflow forecasts have two main uses. First, flood early warning systems (Nevo et al., 2022) where forecast models are integrated with inundation models to predict the depth and extent of the floods (Bhatt et al., 2017). Second, reservoir operation where streamflow forecasts inform releases instead of classic reservoir rule curves (Delaney et al., 2020).

Previous studies have highlighted the benefits of deploying data-driven models, especially those based on deep learning (DL), for accurate and reliable hydrological forecasting (Cheng et al., 2020; Jahangir et al., 2023; Kao et al., 2020; Xiang et al., 2020; Yin et al., 2021). Unlike process-based models, data-driven models act as pattern recognition tools and approximate the underlying function, mapping the exploratory variables (exogenous variables or covariates) to the target (response variable) without considering the physics (e.g., governing differential equations) of the modeled system (Liu et al., 2015; Solomatine and Ostfeld, 2008). It is worth noting that streamflow simulation and forecasting have distinct meanings and should be differentiated. In the machine learning (ML) community, the term forecasting is adopted when future values are obtained using past observations and future forecast signals, while simulations are used to obtain current values of a variable using current (and potentially past) observations (Jahangir et al., 2023) which is in line with the recommendations of Beven and Young (2013). However, in the hydrology community, the terms “prediction” and “forecast” are frequently used interchangeably when applied to rainfall-runoff modeling (Beven and Young, 2013), when in reality the authors of such studies are using simulation models (see, for example, Alizadeh et al., 2021; Apaydin and Sibtain, 2021; Lv et al., 2020). Hereafter, the term “forecasting” is only adopted

when referring to predicting future values, while “prediction” is used for both forecasting and simulation.

This research aims to introduce a novel method for probabilistic hydrological forecasting, namely, conditional variational auto-encoder (CVAE). CVAE is based on encoder-decoder (ED) DL models. Since the focus here is on introducing CVAE for hydrological forecasting and its comparison against DL benchmarks, details regarding rainfall-runoff modeling through process-based and data-driven models are not provided. The interested reader may refer to Knoben et al. (2019), Kratzert et al. (2019) and Zhou et al. (2019) for details regarding recent improvements in process-based and data-driven rainfall-runoff modeling.

Due to the uncertainty inherent in hydrological variables (e.g., precipitation; Kelly and Krzysztofowicz, 2000), it is important to develop probabilistic forecast models when working with these datasets. These models produce not only point forecasts of the variable of interest, such as streamflow, but also indicate the uncertainty associated with the forecast. This uncertainty is typically represented through prediction intervals or probability distributions. Utilizing probabilistic forecasts in hydrological applications (e.g., reservoir operation, flood early warning systems) is especially important when making risk-based decisions that rely on reliable forecasts. By incorporating probabilistic forecasts into risk-based decision-making procedures, decision-makers can minimize costs, reduce economic losses, and improve the resiliency of the systems being modeled (Duque et al., 2023). Furthermore, with probabilistic forecasting, decision-makers can gain a more comprehensive understanding of the potential outcomes of their actions and make better-informed decisions.

In hydrology, there are many approaches for quantifying prediction uncertainty. One popular approach is ensemble-based modeling, where the predictions from each ensemble member are weighted equally (or differently) to arrive at the probability density of the target (e.g., Hauswirth et al., 2023; Montanari and Koutsoyiannis, 2012; Pappenberger et al., 2015; Thiboult et al., 2016; Troin et al., 2021; Velázquez et al., 2009). Other popular approaches include pre- and post-processing methods, such as those based on Bayesian approaches (Diks and Vrugt, 2010; Hao et al., 2022; Höge et al., 2019; Luo and Wood, 2008; Qi et al., 2018). However, an emerging method uses DL models to directly estimate prediction uncertainty through model training (optimization) and/or selection of the DL model structure. This class of DL models are called probabilistic DL models. Instead of generating point predictions (e.g., an estimate of the mean of the target distribution), the conditional probability

distribution of the target (conditioned on the model inputs and parameters) is derived (see Section 3.3 for details). The small number of studies using probabilistic DL for hydrological prediction can be categorized into three groups.

The first group quantifies the uncertainty of the network's weights and biases (parameters) through Bayesian modeling. For example, Lu et al. (2021) adopted Monte-Carlo (MC) dropout (Gal and Ghahramani, 2016) for variational approximation (Jordan et al., 1999) (see Section 3.3) of the posterior distribution of long short-term memory network (LSTM) parameters. Their results suggested that the proposed MC dropout method can effectively estimate the uncertainty of LSTM, improving the comprehension and credibility of the model predictions. Li et al. (2021) explored variational posterior approximation in a Bayesian LSTM framework for uncertainty estimation in process-based hydrological models. They undertook a benchmarking exercise against Bayesian linear regression. Their findings indicated that the proposed method leads to more reliable predictions with narrower prediction intervals. Ghobadi and Kang (2022) utilized variational posterior approximation to estimate LSTM parameters for multi-step ahead streamflow forecasting. Their findings demonstrated that direct variational posterior approximation is more effective for multi-step ahead probabilistic streamflow forecasting than MC dropout-based variational approximation.

The second group of studies concentrated on probabilistic output layers and derived characteristics of the target distribution rather than estimating the mean. Klotz et al. (2022) employed mixed probability distribution derivation (mixture density networks) via LSTM models for rainfall-runoff modeling. The authors tested their framework on 531 catchments in the contiguous United States, employing the Catchment Attributes and MEteorolgy for Large-sample Studies (CAMELS) dataset (Addor et al., 2017). Their results revealed that distributional predictions are reliable and accurate, leading to robust single-point estimates. Nevo et al. (2022) used a start-to-end ML pipeline for flood forecasting, where water level forecasts obtained through probabilistic DL models were akin to the model proposed by Klotz et al. (2022). These forecasts were subsequently utilized as inputs to a flood inundation model for forecasting flood depth and extent.

The third group includes a new type of probabilistic DL model trained using the quantile loss function instead of a traditional loss/objective function, such as mean squared error (MSE). Jahangir et al., (2023) investigated quantile-based ED models coupled with wavelet decomposition for multi-step

ahead runoff forecasting. The newly proposed quantile-based ED models resulted in highly reliable forecasts even when forecast accuracy was poor.

In an ED framework, the input data are encoded through an encoder model into a vector for capturing the impact of local factors. Afterwards, the local factors are mapped to the target by decoding the encoded vector through a decoder model (Apaydin and Sibtain, 2021). The ED structure can extract temporal associations among time-series (Cui et al., 2022). Therefore, the ED configuration can better model non-linear relationships between time-series compared to single network configurations (Bian et al., 2020). The latter is of great importance in hydrological forecasting due to the complex properties of hydrological processes. One of the main advantages of an ED framework compared to other ML models is that there is no restriction on designing the encoder or decoder. Thus, the encoder and decoder can be any type of neural network (recurrent, convolutional, or dense). The latter results in a multitude of choices for designing the ED components. Therefore, ED frameworks offer a flexible approach to obtaining accurate hydrological forecasts.

The present research introduces a new approach to probabilistic streamflow forecasting, utilizing the generative DL (GDL) paradigm (additional details provided in the following subsection). While GDL has been recently applied to tasks such as renewable energy forecasting (Dumas et al., 2022), this study is the first, to the best of the authors' knowledge, to explore the potential of GDL for hydrological forecasting. The following subsection presents a brief overview of the limited studies utilizing GDL for forecasting. This overview describes the merits of GDL compared to the existing approaches mentioned above.

3.1.1 Generative Deep Learning

GDL refers to a set of methods that use deep neural networks to learn and represent the probability distribution of the target (Dumas et al., 2022). Although generative modeling originated in the 1980s, the field has recently seen a surge in interest due to the availability of large, open-access datasets, significant advancements in DL architectures, and the accessibility and prevalence of graphical processing units (GPUs) for deployment of such models (Dumas et al., 2022). In essence, a generative model is a probabilistic model that characterizes the process by which a dataset is generated. The ultimate aim is to utilize the generative model to produce new data that conforms to the original data (Fanfarillo et al., 2021). These models have been adopted in different domains, such as computer vision (e.g., pose estimation), natural language processing (e.g., chatbots), and medicine (e.g., drug synthesis)

(Bond-Taylor et al., 2022). Generative models offer the advantage of flexible neural network architecture design, permitting the use of diverse structures, such as dense (e.g., multi-layer perceptron, MLP), convolutional, and recurrent (e.g., LSTM) networks. This is due to their ED architecture. This structure can effectively model nonlinear relationships between time-series because it can extract temporal associations among them. Compared to single network configurations, ED configurations are better suited for modeling such relationships, as demonstrated by Bian et al. (2020) and Cui et al. (2022).

Various types of GDL models exist, as extensively covered in Bond-Taylor et al. (2022). However, among the most well-known are the variational auto-encoder (VAE), normalizing flow (NF), and generative adversarial network (GAN). Each of these methods has its strengths and weaknesses. For example, GAN models are challenging to train and often result in over-fitting, while it is difficult to optimize the hyperparameters of VAE models (see Bond-Taylor et al. (2022) for a detailed review). This research focuses on developing CVAE, a specific configuration of VAE, for producing medium-range (one-seven day(s) ahead) probabilistic streamflow forecasts. Although there are no examples of VAE (or, more specifically, CVAE) being applied for hydrological forecasting, it has been shown to be an effective method for applications in related domains, such as renewable energy forecasting (Dumas et al., 2022; Huang et al., 2022), ozone forecasting (Cheng et al., 2022), precipitation downscaling (Harris et al., 2022), data assimilation in groundwater models (Bao et al., 2022), and designing water remediation strategies (Kang et al., 2021). In the context of hydrological prediction, VAE-based models have also been adopted as surrogate models, a more efficient alternative than solving shallow water equations for predicting river flows (Forghani et al., 2021, 2022a). The interested reader is referred to (Forghani et al., 2022b) for novel applications of VAE in hydrology.

A notable benefit of GDL models, in contrast to other probabilistic DL models, is their capacity to generate an ensemble of predictive distributions of the target. In other words, instead of producing a single predictive distribution of the target, such as the approaches in (Klotz et al., 2022), GDL models can generate multiple predictive distributions of the target. This leads to improved generalization and a more accurate handling of the uncertainty linked with model configuration. Of note, this is different than ensemble prediction (where multiple models are produced), as only one model is trained and deployed. It is also different than Bayesian DL (approximating the distribution of the network weights) since the network parameters in GDL models are deterministic. Hence, GDL models are more efficient

in training and deployment compared to the Bayesian methods. The following subsection highlights the main contributions of this research and outlines the rest of this chapter.

3.1.2 Contributions and Outline

In this research, a novel CVAE model is developed, evaluated, and applied for probabilistic multi-step ahead streamflow forecasting. To achieve this goal, CVAE models are developed and compared to state-of-the-art quantile-based ED models. The effectiveness of the models is evaluated using 75 basins in Canada selected from the Canadian hydrometeorological watershed database (CANOPEX; Arsenault et al., 2016). An alternative configuration of CVAE, the quantile-based CVAE (QCVAE), is also introduced.

While CVAE outputs the conditional probability distribution of the target, the QCVAE outputs specific quantiles of the conditional probability distribution of the target. Various deterministic and probabilistic metrics are used to evaluate the CVAE against its benchmarks (ED and QCVAE). The main contributions of this study are summarized below:

1. This is the first investigation to explore CVAE models' efficacy for probabilistic multi-step ahead streamflow forecasting. Unlike prior research (e.g., Dumas et al., 2022), neural network structures that have proven useful for streamflow forecasting (i.e., LSTM; Sabzipour et al., 2023)) are used for constructing the encoder-decoder components. The suggested models are flexible, sequence-to-sequence models capable of simultaneously producing all forecast lead times without depending on any future data, such as meteorological (forcing) forecasts (see Section 3.3 for details).
2. The current research is classified as an ad hoc benchmarking and model intercomparison study (Klotz et al., 2022) for probabilistic streamflow forecasting (using probabilistic DL models) since a standard dataset is adopted (i.e., CANOPEX), models are developed, evaluated, and applied based on common practices in the ML community (e.g., data splitting, standardization), and the results are disseminated to the public (Nearing et al., 2018).
3. Unlike previous work involving CVAE (e.g., Dumas et al., 2022), three distinct types of probabilistic models are designed, each is subject to hyperparameter tuning using Bayesian optimization (20 different models assessed) and trained and tested for 75 basins. Additionally, two sets of model inputs (with and without past streamflow observations) are investigated,

resulting in training more than 9000 forecast models (a combination of three models, 20 alternative hyperparameter sets, 75 basins, and two input sets).

The rest of this chapter is structured as follows. Section 3.2 provides a description of the CANOPEX dataset and the key statistical parameters of the selected basins. Section 3.3 outlines the methodology used for developing the GDL models. Afterwards, Section 3.4 presents and analyzes the results of the study. Finally, Section 3.5 concludes this chapter by discussing the limitations of the work, providing final remarks, and highlighting possible future research directions.

3.2 Dataset

In this study, the CANOPEX dataset was used to develop and evaluate the models. This dataset includes daily meteorological and streamflow data for 698 basins in Canada (Arsenault et al., 2016). The meteorological data are basin-averaged, including precipitation as well as maximum (max) and minimum (min) air temperature. Two types of meteorological data are available in the dataset. The first is based on weather station data and the second is based on a gridded climate product. Like the study conducted by Jahangir and Quilty (2022), the daily 10 km gridded data (see Hutchinson et al., 2009) was adopted as it contained no missing data.

Of the available 698 basins, 75 basins were selected for model development and evaluation (see Appendix B) based on two criteria. The first criterion required the basins to have more than 4000 data samples (approximately 10 years of training data, see below for details), a reasonable number of records for training DL models. The second criterion required no missing streamflow data. Thus, complete records were available for the target (streamflow) and inputs (precipitation, max and min air temperature) for all basins. Afterwards, the data for each basin was divided into three sets for model development and evaluation: training (80 %), validation (10 %), and testing (10 %). The training set was used for input normalization and parameter optimization, while the validation set was used for hyperparameter tuning. The test set was exclusively used for evaluating the results.

Key statistical parameters (min, median, max, and standard deviation (std)) were estimated for each variable used in the modeling procedure across all 75 basins (Table 3-1). Each of the basins' average meteorological and streamflow observations' values are reported in the Appendix B. A preliminary examination of the data confirmed that, in most basins, the streamflow time-series exhibited high autocorrelation and showed a strong influence due to snowmelt in the Spring. The selected basins were distributed across Canada, with a greater concentration in British Columbia, and Newfoundland and

Labrador, relative to the other provinces. The selected basins include both small (292 km²) and extremely large (114, 248 km²) basins (median of 1168.82 km², and a std of 15179.15 km²). The selected basins are depicted in Figure 3-1.

Table 3-1. Statistical parameters of the meteorological variables and streamflow (calculated across the 75 basins).

		Statistical parameters			
		min	median	max	std
Training (80%)	Precipitation [mmday ⁻¹]	1.20	2.72	4.09	0.73
	Max. daily temperature [°C]	1.64	7.59	12.02	2.28
	Min. daily temperature [°C]	-8.74	-2.67	2.48	2.40
	Streamflow [mmday ⁻¹]	1.73	25.11	820.62	152.34
Validation (10%)	Precipitation [mmday ⁻¹]	1.12	2.51	4.01	0.73
	Max. daily temperature [°C]	2.14	8.09	12.36	2.45
	Min. daily temperature [°C]	-8.36	-2.02	3.65	2.55
	Streamflow [mmday ⁻¹]	1.82	24.70	833.49	154.82
Test (10%)	Precipitation [mmday ⁻¹]	1.16	2.62	3.97	0.75
	Max. daily temperature [°C]	2.13	7.99	12.83	2.38
	Min. daily temperature [°C]	-8.09	-1.77	3.77	2.69
	Streamflow [mmday ⁻¹]	0.80	25.89	804.26	147.37

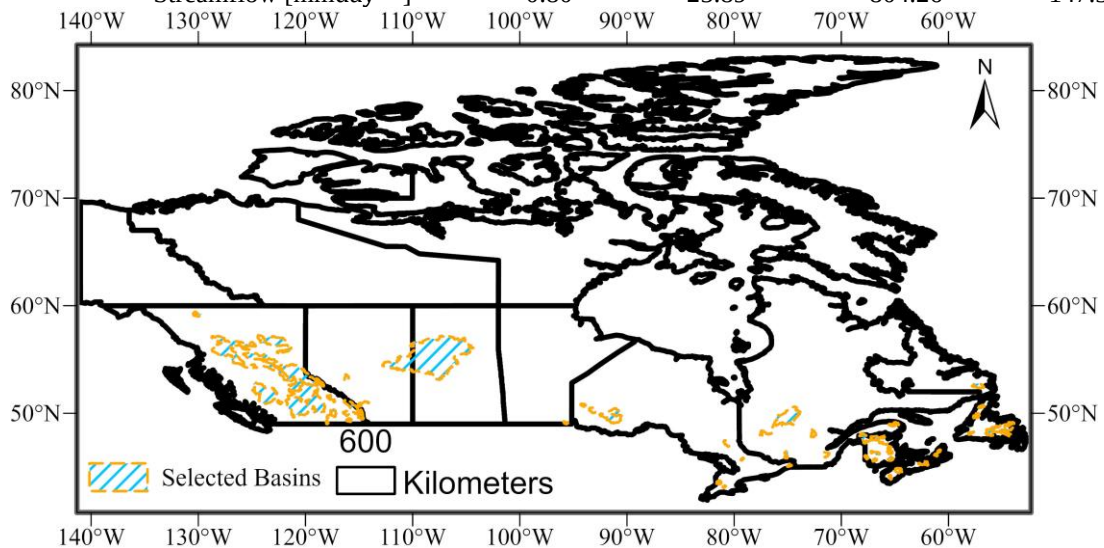


Figure 3-1. The 75 selected basins across Canada.

3.3 Methods

This section describes the methodology for developing probabilistic forecast models and introduces the evaluation criteria. The study considered three types of forecast models: ED (quantile-based encoder-decoder), CVAE (conditional variational auto-encoder), and QCVAE (a quantile-based variant of CVAE). The three models were applied to all 75 basins, with two different input variable sets denoted by superscripts Q and NQ, with and without previous streamflow observations as inputs (see Section 3.3.5 for details regarding the models' input), respectively. For example, the QCVAE^{NQ} model is the quantile-based CVAE model with no past streamflow observations. All forecast models were used for seven-day ahead forecasting, simultaneously producing forecasts for lead times one-seven.

A fourth model was developed exclusively for simulation purposes, aimed at predicting streamflow at time step t . This model was used as a benchmark for the forecast models. For this purpose, a LSTM model similar to the Gaussian mixture model proposed by Klotz et al. (2022) was adopted. The simulation models consisted of an LSTM block, a dense layer, and a Gaussian component as the output layer. As the model was considered solely for simulation, no past streamflow observations were used as input to this model (only past meteorological data was used as input). i.e., a probabilistic rainfall-runoff model was developed.

Quantile-based models estimate the lower quantile (0.05, q_L), median (q_M), and the upper quantile (0.95, q_U) for all forecast steps simultaneously. In contrast, CVAE and the simulation model estimate the joint probability distribution of the forecasted steps (the joint probability distribution of all lead times). The general structure of the developed models is illustrated in Figure 3-2. Since the designed models were based on coupling a LSTM with a dense network (LSTM-DN), a brief introduction to the LSTM architecture is provided in the subsection below before introducing the main models.

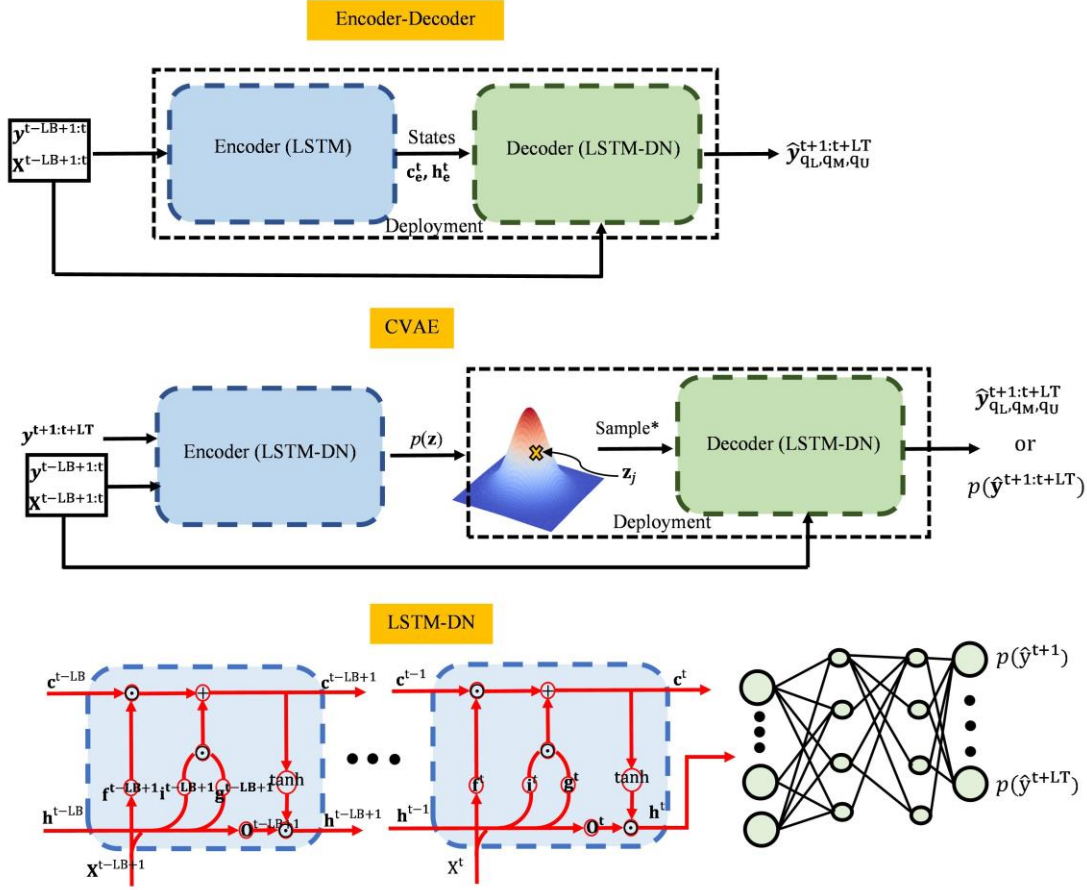


Figure 3-2. General ED, CVAE, and LSTM-DN structure scheme. $\mathbf{X}$ is the input array to the models, while $\mathbf{y}$ indicates the target (streamflow). For high quality version of this figure, the reader should refer to the published paper.

3.3.1 Long Short-Term Memory Network

RNNs, and especially LSTMs, have demonstrated impressive performance in outperforming classical machine learning models, such as multi-layer perceptron and support vector machines, for hydrological prediction (Hu et al., 2018; Kratzert et al., 2018; Le et al., 2019). RNNs were designed to handle temporal dynamics using feedback connections (Elman, 1990), which makes them suitable for analyzing nonlinear hydrological series (Coulibaly and Baldwin, 2005; Ni et al., 2020). LSTMs introduced in the late 1990s by Hochreiter and Schmidhuber (1997) sought to address the gradient exploding/vanishing problem encountered by RNNs when processing long data sequences. The

gradient exploding/vanishing problem arises due to the vanishing gradient or the exponential growth of gradients that occurs during backpropagation through time, which makes it difficult for RNNs to learn long-term dependencies. LSTMs address this problem through internal memory cells and gating mechanisms, allowing them to remember or forget information over extended periods selectively.

The formulation of a LSTM cell associated with timestep T is as follows (Jahangir et al., 2023; Kratzert et al., 2019):

$$\mathbf{i}^T = \text{sigmoid}(\mathbf{W}_i \mathbf{X}^T + \mathbf{U}_i \mathbf{h}^{T-1} + \mathbf{b}_i) \quad \text{Eq 3-1}$$

$$\mathbf{f}^T = \text{sigmoid}(\mathbf{W}_f \mathbf{X}^T + \mathbf{U}_f \mathbf{h}^{T-1} + \mathbf{b}_f) \quad \text{Eq 3-2}$$

$$\mathbf{g}^T = \tanh(\mathbf{W}_g \mathbf{X}^T + \mathbf{U}_g \mathbf{h}^{T-1} + \mathbf{b}_g) \quad \text{Eq 3-3}$$

$$\mathbf{o}^T = \text{sigmoid}(\mathbf{W}_o \mathbf{X}^T + \mathbf{U}_o \mathbf{h}^{T-1} + \mathbf{b}_o) \quad \text{Eq 3-4}$$

$$\mathbf{c}^T = \mathbf{f}^T \odot \mathbf{c}^{T-1} + \mathbf{i}^T \odot \mathbf{g}^T \quad \text{Eq 3-5}$$

$$\mathbf{h}^T = \mathbf{o}^T \odot \tanh(\mathbf{c}^T) \quad \text{Eq 3-6}$$

where $\text{sigmoid}(x) = 1/e^{-x}$, $\tanh(x) = (e^{2x} - 1)/(e^{2x} + 1)$, and $\mathbf{g}^T$, $\mathbf{h}^T$, $\mathbf{c}^T$ are the cell input, recurrent state, and cell state at timestep T , respectively. The $\odot$ operator indicates the element-wise multiplication. The output gate ($\mathbf{o}^t$) controls which saved information is outputted and a loop is used for feeding the previous cell states ($\mathbf{c}^{t-1}$) and recurrent states ($\mathbf{h}^{t-1}$) into the network.

In the LSTM-DN configuration, the output from the last iteration (hidden state vector, $\mathbf{h}^t$) is fed to a DN with two hidden layers. In the CVAE model, the DN is used for estimating the joint probability distribution of the target (joint probability density function of the forecasts across the seven lead times), $p(\hat{\mathbf{y}}^{t+1:t+LT})$. In this research, $p(\hat{\mathbf{y}}^{t+1:t+LT})$ was obtained by assuming independent normal (Gaussian) distributions for each forecast step:

$$p(\hat{\mathbf{y}}^{t+1:t+LT}) = \prod_{j=1}^{LT} p(\hat{y}^{t+j} | \mu^{t+j}, \sigma^{t+j}) \quad \text{Eq 3-7}$$

$$p(\hat{y}^{t+j} | \mu^{t+j}, \sigma^{t+j}) \sim \mathcal{N}(\mu^{t+j}, \sigma^{t+j}) \quad \text{Eq 3-8}$$

where μ and σ are the mean and std of a Gaussian distribution ($\mathcal{N}$). For simplicity, $p(\hat{y}^{t+j}|\mu^{t+j}, \sigma^{t+j})$ is hereafter represented by $p(\hat{y}^{t+j})$. For simulation, the LSTM-DN was utilized for estimating the target distribution at time step t ($\mathcal{N}(\mu^t, \sigma^t)$) without using past streamflow observations as input.

3.3.2 Quantile-Based Encoder-Decoder

This research explored a modified version of the methodology proposed by Jahangir et al. (2023) for developing ED models. In this configuration, the encoder is a LSTM network that produces two vectors, $\mathbf{h}_e^t$, and $\mathbf{c}_e^t$. The two vectors are then passed to the decoder as the initial states of the decoder's LSTM. In the proposed configuration, the input array $\mathbf{X}^{t-LB+1:t} \in \mathcal{R}^{n \times LB \times m}$ was shared between the encoder and the decoder. The DN used the output of the decoder's LSTM, $\mathbf{h}_d^t$, to obtain the three designated quantiles (q_L, q_M, q_U) of the target distribution ($\hat{\mathbf{y}}_{q_L, q_M, q_U}^{t+1:t+LT}$). Unlike the study by Jahangir et al. (2023) where three different models were optimized for obtaining $\hat{\mathbf{y}}_{q_L, q_M, q_U}^{t+1:t+LT}$, the last layer of the DN was split into three different branches (each associated with one quantile of interest), and every branch was regularized accordingly (Wang et al., 2019):

$$L_{q_i}^t = \begin{cases} \sum_{j=1}^{LT} (1 - q_i) (\hat{y}_{q_i}^{t+j} - y^{t+j}) & \hat{y}_{q_i}^{t+j} \geq y^{t+j} \\ \sum_{j=1}^{LT} (q_i) (y^{t+j} - \hat{y}_{q_i}^{t+j}) & \hat{y}_{q_i}^{t+j} < y^{t+j} \end{cases} \quad \text{Eq 3-9}$$

where $L_{q_i}^t$ is the pinball loss associated with the q_i th quantile at time step t , $\hat{y}_k^{t+j}$ and y_k^{t+j} are the forecasted and the target streamflow (lead time $t+j$) associated with the q_i th quantile, respectively.

3.3.3 Conditional Variational Auto-Encoder

CVAE is a deep latent variable model comprising two separate networks jointly trained (Dumas et al., 2022; Salazar et al., 2022). The training and forecasting stages within a CVAE model follow different procedures and should be distinguished.

Training CVAE models follows multiple steps. In the first step, the encoder maps the input (target and all other exogenous variables to the forecast models) to a stochastic latent variable $\mathbf{z}^t \in \mathcal{R}^d$. Consider $\mathbf{x}^t = (\mathbf{X}^{t-LB+1:t}, \mathbf{y}^{t+1:t+LT})$ and the encoder network (LSTM-DN) being parametrized by Φ . Then, the encoder $p_\Phi(\mathbf{z}^t|\mathbf{x}^t)$ acts as a surrogate distribution and is designed to approximate the true

posterior of the latent variable $\mathbf{z}^t$ given by $p(\mathbf{z}^t)$. The true posterior of $\mathbf{z}^t$ is predefined and was set as a multivariate normal distribution with zero mean and unit covariance matrix ($\mathbf{I}$).

In the next step, consider the decoder (LSTM-DN) parametrized by $\boldsymbol{\phi}$. Then, $p_{\phi}(\hat{\mathbf{y}}^{t+1:t+LT}|\mathbf{z}^t, \mathbf{X}^{t-LB+1:t})$ approximates the target distribution, $p(\mathbf{y}^{t+1:t+LT})$. In other words, the decoder maps the latent variable ($\mathbf{z}^t$) alongside the exogenous variables ($\mathbf{X}^{t-LB+1:t}$) to the target ($\mathbf{y}^{t+1:t+LT}$). The two networks are jointly optimized by maximizing the evidence lower bound (ELBO; Blei et al., 2017; Kingma and Welling, 2022):

$$L_{\Phi, \phi}^t = -D_{KL}(p_{\Phi}(\mathbf{z}^t|\boldsymbol{\chi}^t)|p(\mathbf{z}^t)) + \mathbb{E}_{p_{\Phi}}(\log(p_{\phi}(\hat{\mathbf{y}}^{t+1:t+LT}|\mathbf{z}^t, \mathbf{X}^{t-LB+1:t}))) \quad \text{Eq 3-10}$$

where $\mathbb{E}_{p_{\Phi}}$ is the expectation of the log-likelihood calculated concerning the encoder weights and D_{KL} is the Kullback-Leibler divergence (Kullback and Leibler, 1951). Of note, the encoder ($p_{\Phi}(\mathbf{z}|\boldsymbol{\chi})$) and decoder ($p_{\phi}(\mathbf{y}|\mathbf{z}, \mathbf{x})$) are both neural networks of choice, where LSTM-DN were selected in this study.

After the training is finished, the optimized decoder can be utilized for forecasting. In the forecasting stage, for each time step, random vectors (determined by the pre-defined number of ensemble members) were sampled from the pre-defined distribution of $\mathbf{z}$ ($p(\mathbf{z})$). The sampled vectors were fed to the decoder alongside the exogenous variables for obtaining the forecasts (see Figure 3-2). The latter produces an ensemble of predictive distributions instead of obtaining a single predictive distribution (see Eq 3-8). In the current study, 100 sample vectors were fed to the decoder, and the median predictive distribution was derived. In other words, 100 predictive distributions ($p(\hat{\mathbf{y}}^{t+1:t+LT})$) were obtained through sampling of the decoder, and the median of the predictive distribution parameters was used for the final predictive distribution.

3.3.4 Quantile-Based Conditional Variational Auto-Encoder (QCVAE)

QCVAE is the non-parametric counterpart of the CVAE, where the main difference is how the network weights are optimized. Like ED, the last DN layer was split into three branches and each branch was regularized (see Eq 3-9). Details regarding the network optimization function are provided in Appendix B.

3.3.5 Model Development

For developing the models, two different input variable sets were considered. The first set used only the meteorological observations (precipitation, max and min air temperature) up to and including

current timestep t to forecast the target, seven day-ahead streamflow ($y^{t+1:t+7}$). The second model input configuration was akin to the first, with the addition of past streamflow observations (up to and including current timestep t). Since the CANOPEX dataset does not contain any meteorological forecast data, consistent with similar studies (Han et al., 2021; Jahangir et al., 2023; Kao et al., 2020), the models did not consider meteorological forecasts. Nevertheless, the proposed models have a flexible structure and can easily be adapted to incorporate meteorological forecasts (Jahangir et al., 2023)

For harnessing the power of the recurrent neural networks (i.e., the LSTM blocks embedded in the models, see Section 3.1), a long look-back (LB) period of 365 days was used. This allowed one year of historical observations to be fed into the DL models. This LB period was selected based on previous successful implementations of LSTM models for hydrological prediction (e.g., Kratzert et al., 2018, 2019). Searching for an optimal LB was beyond the scope of the current study due to the high computational requirements. Before training, both the input and target were z-normalized using the training set statistics to make the convergence to optimal parameters (network weights and biases) more efficient. Consequently, the model outputs were back-transformed when evaluating model performance.

Models' hyperparameters (for all basins, input variable sets, and models) were tuned through Bayesian optimization. Bayesian optimization, derived from Bayes theorem, is an effective method for solving computationally demanding functions that do not have closed form solutions (Brochu et al., 2010). Thus, Bayesian optimization can be leveraged for optimizing the structure of DNNs through hyperparameter selection. The objective function (i.e., the pinball loss function) is approximated by a Gaussian process (Seeger, 2004). By considering an acquisition function, where high values of the acquisition function are associated with high values of the objective function, the maximum value of the objective function is obtained through an iterative process of maximizing the acquisition function at each step of Bayesian optimization (see Wu et al. (2019) for details). The optimization process included 20 iterations, with each iteration training the neural networks for 20 epochs (optimizing three model types, two input types, 75 basins, and 20 iterations resulting in 9000 models). The hyperparameters include the LSTM hidden state size (ranging from 8 to 256), and DN layer activation function type (exponential linear unit, linear). DN layer size was set based on the output layer's size. For example, for CVAE, the DN size was set to 14, which is due to estimating normal distribution parameters for seven forecast steps. The best hyperparameter set was selected based on the model performance for the validation set, and the corresponding model was optimized for up to 200 epochs.

Thus, there was a final optimization stage that included three model types, two input types, and 75 basins, resulting in 450 models.

The hyperparameters selected across the 75 basins and associated with the generative models (CVAE and QCVAE) considering past streamflow observations are shown in Appendix B. The hyperparameters include the LSTM layers' size (for both encoder and decoder) and the DNs' (encoder) activation function type. The analysis reveals that, in the case of CVAE, the linear activation function was notably favored over the exponential linear unit activation function, especially for the decoder. Conversely, for QCVAE, the exponential linear unit was more commonly selected as the activation function. Additionally, it was observed that for CVAE, larger LSTM networks tended to be selected for the decoder as compared to the encoder, while for QCVAE, the encoder LSTM networks were larger than the decoder networks. The difference between the hyperparameters selected for these two types of models is a significant finding, as it indicates that the selection of the objective function of DL models (e.g., maximizing ELBO) exerts a substantial influence on the models' structure. This opens an intriguing avenue for further research, one that has not yet been explored extensively in the field of hydrology. Thus, future studies focused on the development of deep learning-based hydrological forecasting models may benefit from a detailed exploration of different objective functions for optimizing CVAE models.

The adaptive moment estimation (Adam) optimization method (Kingma and Welling, 2022) was used for parameter optimization, and a dynamic learning rate with an initial value of $1e-4$ and a decreasing factor of 0.5 was employed for all models. The decreasing factor was applied every three epochs unless validation loss decreased. Furthermore, to avoid overfitting, if the validation loss did not decrease for 10 consecutive epochs, training was terminated. Custom scripts in Python 3.9 using the TensorFlow 2.4.1 library with the Keras API were used to develop the models. All calculations were performed on the Graham cluster provided by Alliance Canada (<https://docs.alliancecan.ca/wiki/Graham>) using a single GPU computational node (NVIDIA V100 Volta) and 32 GB of RAM.

3.3.6 Forecast Evaluation

Common deterministic and probabilistic metrics from the hydrological modeling literature were adopted for evaluating the forecasts. All reported metrics are associated with the test set, the data partition not used for model training or hyperparameter tuning. All forecast steps (one to seven) were

compared against the same target using a rolling window to ensure the performance of the models across different forecast steps was comparable.

The deterministic metrics include the normalized root mean square (nRMSE) and modified Kling–Gupta efficiency (KGE; Gupta et al., 2009; Kling et al., 2012):

$$\text{nRMSE} = \sqrt{1/n \sum_{k=1}^n (\hat{y}^k - y^k)^2} / \mu_y \quad \text{Eq 3-11}$$

$$\text{KGE} = 1 - \sqrt{(\rho - 1)^2 + \left(\frac{\sigma_{\hat{y}}}{\sigma_y} - 1\right)^2 + \left(\frac{\mu_{\hat{y}}}{\mu_y} - 1\right)^2} \quad \text{Eq 3-12}$$

where μ and σ are the mean and std deviation operators, respectively. The parameter n indicates the total number of samples and ρ is the linear correlation (Pearson correlation) between forecasts and observations. The probabilistic metrics include the prediction interval coverage probability (reliability; PICP) and mean prediction interval width (MPIW) were adopted (Gneiting and Raftery, 2007):

$$\text{PICP} = \frac{1}{n} \sum_{k=1}^n U_k, U_k = \begin{cases} 0 & y^k \notin [\hat{y}_{q_L}^k, \hat{y}_{q_U}^k] \\ 1 & y^k \in [\hat{y}_{q_L}^k, \hat{y}_{q_U}^k] \end{cases} \quad \text{Eq 3-13}$$

$$\text{MPIW} = \frac{1}{n} \sum_{k=1}^n (\hat{y}_{q_U}^k - \hat{y}_{q_L}^k) \quad \text{Eq 3-14}$$

where q_L , and q_U were set as 0.05, and 0.95, respectively. Based on the defined 90% confidence level ($q_U - q_L$), a PICP of 90% with the smallest MPIW possible is sought. The forecast quantiles were obtained directly in ED and QCVAE. However, for CVAE and the simulation model, the forecast quantiles were estimated using the predictive distribution.

This study also examined the influence of different features of the input variables on model performance. Different features of the input variables included training sample size as well as the coefficient of variation of precipitation and streamflow. Furthermore, the effect of autocorrelation at three different lags was investigated for streamflow. Specifically, the sum squared autocorrelation value for 15 (ρ_{15}^2), 90 (ρ_{90}^2), and 365 (ρ_{365}^2) lags were computed (please see Papacharalampous et al. (2021) for details) to see if relationships exist between model performance and the abovementioned features.

3.4 Results and Discussion

In this section, the performance of the forecasts is presented, analyzed, and discussed. At first, the deterministic performance of the models is explored, followed by an evaluation of the models' probabilistic performance. Finally, an investigation is undertaken to assess the impact of different features of the input variables on model performance. As outlined in Section 3.2, two input variable sets were used to develop the forecast models. The models incorporating past streamflow observations as input are denoted by the superscript Q (ED^Q , $CVAE^Q$, $QCVAE^Q$), while the models that did not use past streamflow observations are denoted by the superscript NQ (ED^{NQ} , $CVAE^{NQ}$, $QCVAE^{NQ}$). Table 3-2 reports the mean and std of both deterministic and probabilistic metrics. The mean and std for each metric are calculated as follows:

1. Each metric (e.g., KGE) was estimated for each forecast lead time and basin, resulting in seven values for each metric and basin.
2. The average of the seven values for each metric was taken for each basin, resulting in 75 values for each metric.
3. The mean and std of the 75 values for each metric were taken, resulting in Table 3-2.

Table 3-2. Mean and std for each metric averaged across forecast lead times and all 75 basins compared to the mean and std of the simulations (averaged across all 75 basins).

Model	Metric							
	KGE		nRMSE		PICP		MPIW	
	mean	std	mean	std	mean	std	mean	std
Simulation	0.76	0.14	0.61	0.35	90.76	10.66	86.04	156.39
ED^{NQ}	0.69	0.20	0.74	0.44	81.77	7.72	63.79	109.51
ED^Q	0.80	0.20	0.55	0.43	87.01	3.41	35.38	60.25
$CVAE^{NQ}$	0.64	0.21	0.77	0.44	86.60	9.20	76.35	113.81
$CVAE^Q$	0.79	0.21	0.56	0.45	92.05	2.78	42.55	62.56
$QCVAE^{NQ}$	0.69	0.19	0.76	0.46	69.24	14.81	47.76	85.38
$QCVAE^Q$	0.80	0.20	0.57	0.44	83.45	4.90	31.19	49.62

3.4.1 Evaluating Deterministic Performance

The results in Table 3-2 confirm that utilizing past streamflow observations in the forecast models considerably improves the forecast accuracy. In general, incorporating past streamflow observations in the forecast models resulted in more accurate forecasts compared to simulation. It is important to note

that the simulations are based on an approach similar to the one used in (Klotz et al., 2022). This experiment aims to show how the forecast models incorporating previous streamflow can outperform a recent benchmark run in simulation mode. This is an important finding as the simulations are expected to perform better than the forecasts since there is less uncertainty when predicting streamflow in the present versus the future. Thus, it can be said that both the model structure and input variables used in the model are responsible for the strong forecasting performance of the models. The mean KGE of the one-seven day(s) ahead forecasts (see Table 3-2) of ED^Q , $CVAE^Q$, and $QCVAE^Q$ showed 5.26%, 3.95%, and 5.26% relative improvement compared to the simulation models, respectively. In Appendix B relative KGE differences between the forecast models and simulation model are compared across all forecast lead times. The findings affirm that, apart from the CVAE model, the one day-ahead forecasts (one-step ahead) consistently exhibited higher accuracy than the benchmark simulation model, regardless of whether past streamflow observations were incorporated or not. Moreover, when past streamflow observations were introduced into the models, their performance remained superior for forecasts up to four days ahead. In particular, the four ahead forecasts from the ED^Q , $CVAE^Q$, and $QCVAE^Q$ models show a relative improvement of 3.10%, 1.33%, and 2.85%, respectively, over the simulation model.

The performance of the forecast models across all basins and lead times is depicted in Figure 3-3 (where LT denotes a specific forecast lead time). The results confirm that even without considering previous streamflow observations as input, the forecast models had acceptable performance, with a minimum median KGE of 0.65 being obtained by $CVAE^{NQ}$ (for LT 7). It is observed that both quantile-based models, ED and QCVAE, outperformed CVAE in terms of point accuracy across all lead times.

Complementary to the results shown in Table 3-2, the average KGE scores (across LT=1:7) were compared to simulation KGE values and are illustrated in Figure 3-4. Each point in Figure 3-4 represents a basin, with the y-axis indicating the difference between forecast KGE (averaged for all LTs) and simulation KGE. Each subfigure is divided into four sections. The upper right section (green background), is associated with accurate forecasts ($KGE > 0.6$; Nevo et al., 2022) and KGE improvement in forecasting, while the lower left section (highlighted with red) is associated with inaccurate forecasts in which the simulations had higher accuracy. In each section, the number of basins in that section and the average percent change in KGE is reported. For example, the average KGE across the forecasts for lead times one-seven was higher than 0.6 in 47 basins when using $CVAE^Q$.

Furthermore, CVAE^Q led to an average improvement in KGE of 0.13 when compared to the simulation model.

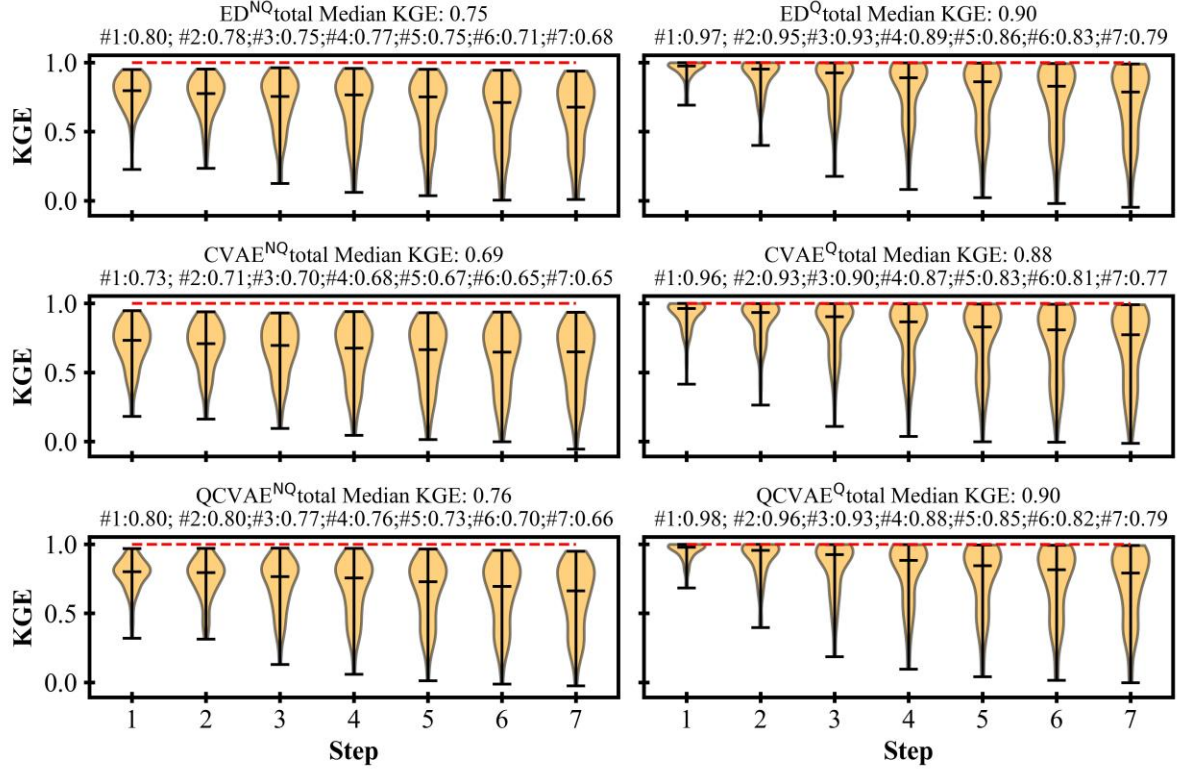


Figure 3-3. KGE violin plot for all lead times. The overall median (across all lead times) and median values for each lead time (#LT) are mentioned above each subplot. The dashed line indicates the desired KGE value (one).

The above results support the need for more complex models and training schemes, as presented in this study, to improve hydrological prediction. Notably, the simulation model, a LSTM-DN with a probabilistic output layer akin to the one used by Klotz et al. (2022) and Nevo et al. (2022), used the same input as the ED^{NQ} , CVAE^{NQ} , and QCVAE^{NQ} models (i.e., only meteorological data). Despite using the same inputs, the simulation model yielded lower accuracy than ED^{NQ} , CVAE^{NQ} , and QCVAE^{NQ} in 30.67% (23/75), 20% (15/75), and 30.67% (23/75) of basins, respectively.

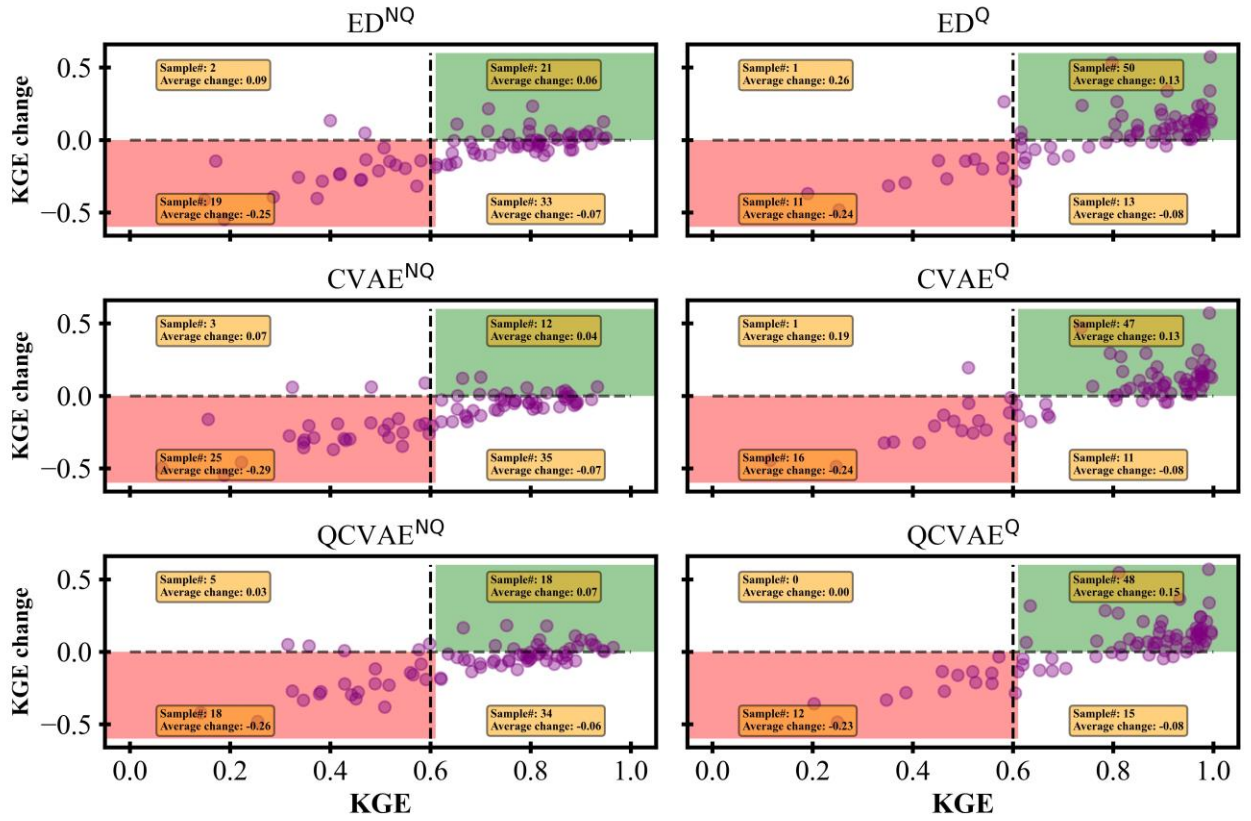


Figure 3-4. Comparison of average forecast KGE with simulation KGE.

3.4.2 Evaluating Probabilistic Performance

The results in Table 3-2 indicate that considering the 90 % confidence level, the CVAE model resulted in probabilistic forecasts that were more reliable than ED and QCVAE. Similar to the deterministic metrics, incorporating past streamflow observations into the models improved the models' performance. The latter finding is supported by Figure 3-5, where larger PICP values with smaller spread were obtained for ED^Q, CVAE^Q, and QCVAE^Q. The fact that CVAE generally produced better probabilistic forecasts but had lower point accuracy compared to the other two model types stems from its training procedure. In contrast to ED and QCVAE, where lower, median, and upper forecast quantiles were obtained directly and simultaneously, CVAE derives a predictive distribution for the decoder. In CVAE, ELBO was maximized (see Eq 3-10), where the trade-off between generating a custom compressed representation of the target through the encoder and reconstructing it using the decoder was balanced.

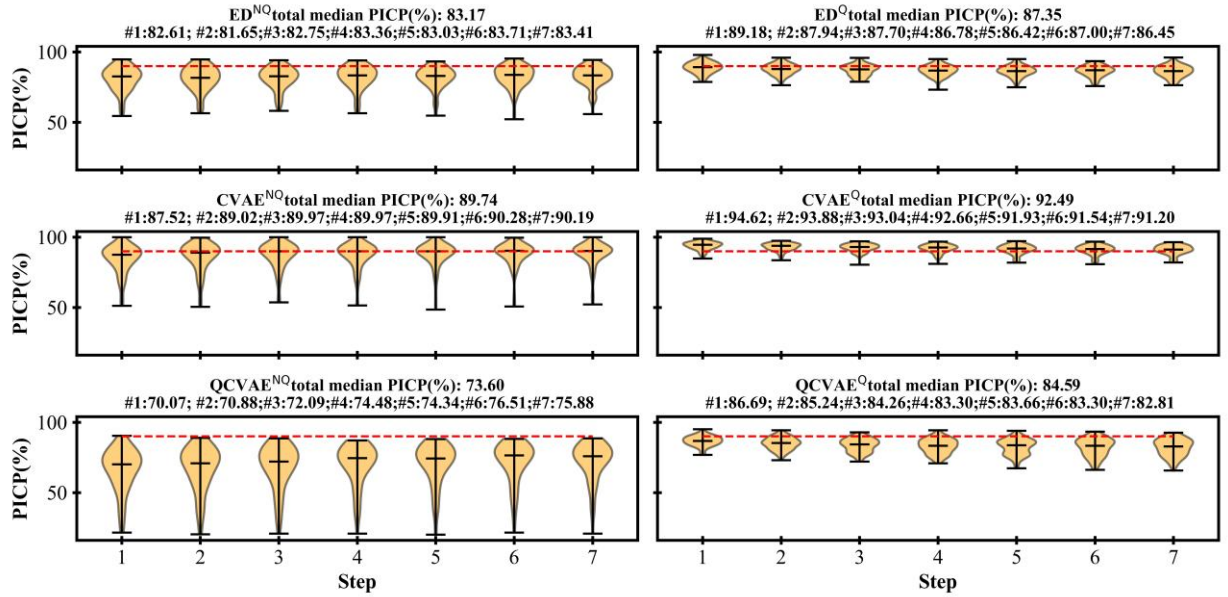


Figure 3-5. PICP violin plot for all lead times. The overall median (across all lead times) and (#LT) median values for each lead time are mentioned above each subplot. The dashed line indicates the desired PICP value (90 %).

To explore the performance of hydrological forecasts from a qualitative point of view, hydrographs are useful. Thus, probabilistic forecasts generated by the DL models at three different lead times (LT = 1, 3, and 7) are presented in Appendix B for a randomly selected basin (specifically, basin 01BJ003). It can be seen that all DL models forecast the target hydrograph reasonably well (i.e., KGE > 0.4 at a seven-day forecast lead time). Although, the DL models that incorporate previous streamflow observations as input are better able to capture peak flows within their prediction intervals. As expected, the forecast accuracy decreases as a function of lead time. However, CVAEQ is the only model to maintain high reliability across all lead times (see Figure 3-5).

3.4.3 Impact of Different Features of the Input Variables on Model Performance

The impact of different features of the input variables on model performance for CVAEQ and EDQ are shown in Figure 3-6, and Appendix B, respectively. Each point represents a different basin, and each subplot is related to a different feature. The different features (training sample size, coefficient of variation of precipitation and streamflow, and sum squared autocorrelation of the streamflow time-series) are plotted against the KGE or PICP, which are averaged values across the one-seven day forecast lead times. A linear regression model was fitted to each pair, and the corresponding regression

coefficient (slope coefficient) and the associated p-value is shown. Of note, a p-value<0.05 shows a significant impact considering a 90% confidence level (two-tailed test).

The different features were selected as they were identified to have the most influence on the performance of the models. Training sample size impacted the optimization of the models. The coefficient of variation was selected as the indicator of variability of the main hydrological variables (streamflow and precipitation). Autocorrelation was used to quantify the temporal dependency in the streamflow time-series.

For both CVAE^Q and ED^Q it is understood that the training sample size did not substantially impact the models' performance. This is attributed to the selection criterion of the basins (see Section 3.2), where only basins with more than 4000 samples were included in the study. The models exhibited lower KGE values for basins with high variations coefficients for precipitation and streamflow time-series. Aligned with the findings of Jahangir et al. (2023) and Feng et al. (2020), high accuracy was achieved for basins with high streamflow autocorrelation.

The analysis reveals a clear distinction between CVAE^Q and ED^Q results. For CVAE^Q, there is an inverse relationship between point accuracy and probabilistic performance. The model compensated for inaccurate point estimates by generating larger prediction intervals, resulting in PICP values close to the desired 90% confidence level. While the forecasts produced by ED^Q were more accurate compared to CVAE^Q, they were not as informative uncertainty-wise. This was confirmed by stepwise performance analysis, where for each forecast step, a linear regression line was fitted to the pairs of models KGE and PICP, elucidating any underlying relationship between probabilistic and deterministic performance (see Figure 3-7). For CVAE^Q, except for LT=1, a significant pattern was observed between PICP and KGE. However, for ED^Q (see Appendix B), it is observed that no significant (all p-values larger than 0.05) linear relationship exists between KGE and PICP values. The pattern of PICP invariance for the ED^Q models is also observed for the features. i.e., the resulted PICP was observed to be invariant catchment related properties such as precipitation variability and streamflow autocorrelation. This consistency can be attributed to the ED^Q models producing monotonic prediction intervals.

Furthermore, the autocorrelation analysis demonstrates that after a specific threshold of ρ^2 (~20), no discernable performance difference was observed between the basins. The latter finding is noteworthy, as it supports the findings of Jahangir et al. (2023) that shorter LBs can lead to more

efficient models and that approach the performance (and sometimes outperform) models with longer LBs, especially when past streamflow observations are utilized in the model (see also Sun et al., 2022).

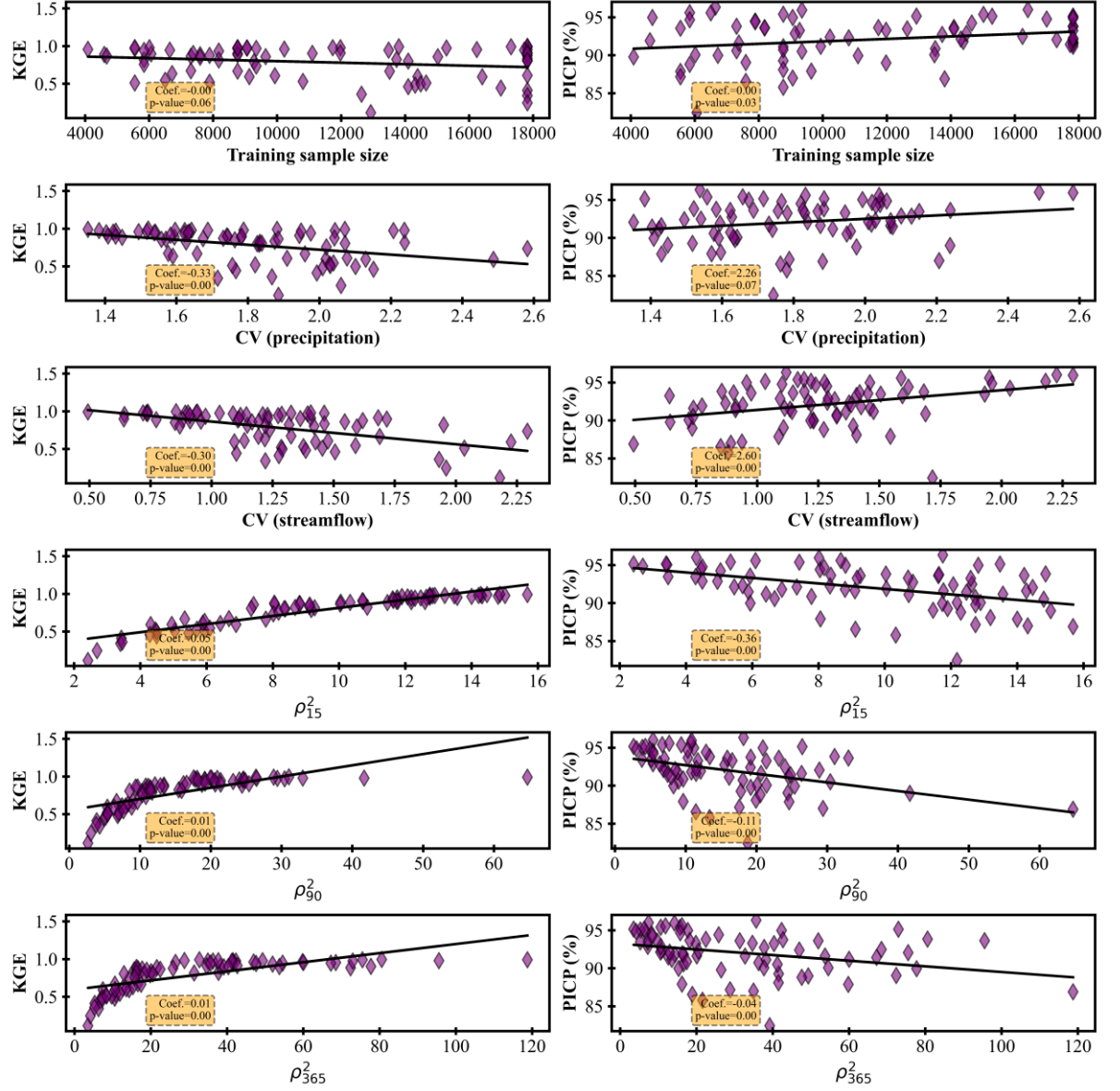


Figure 3-6. Impact of different features of the input variables on the performance of CVAE^Q. KGE and PICP values are averaged across forecast lead times of one-seven days ahead. The fitted lines are based on linear regression where their associated coefficient (slope coefficient) and the associated p-values are shown in the boxes.

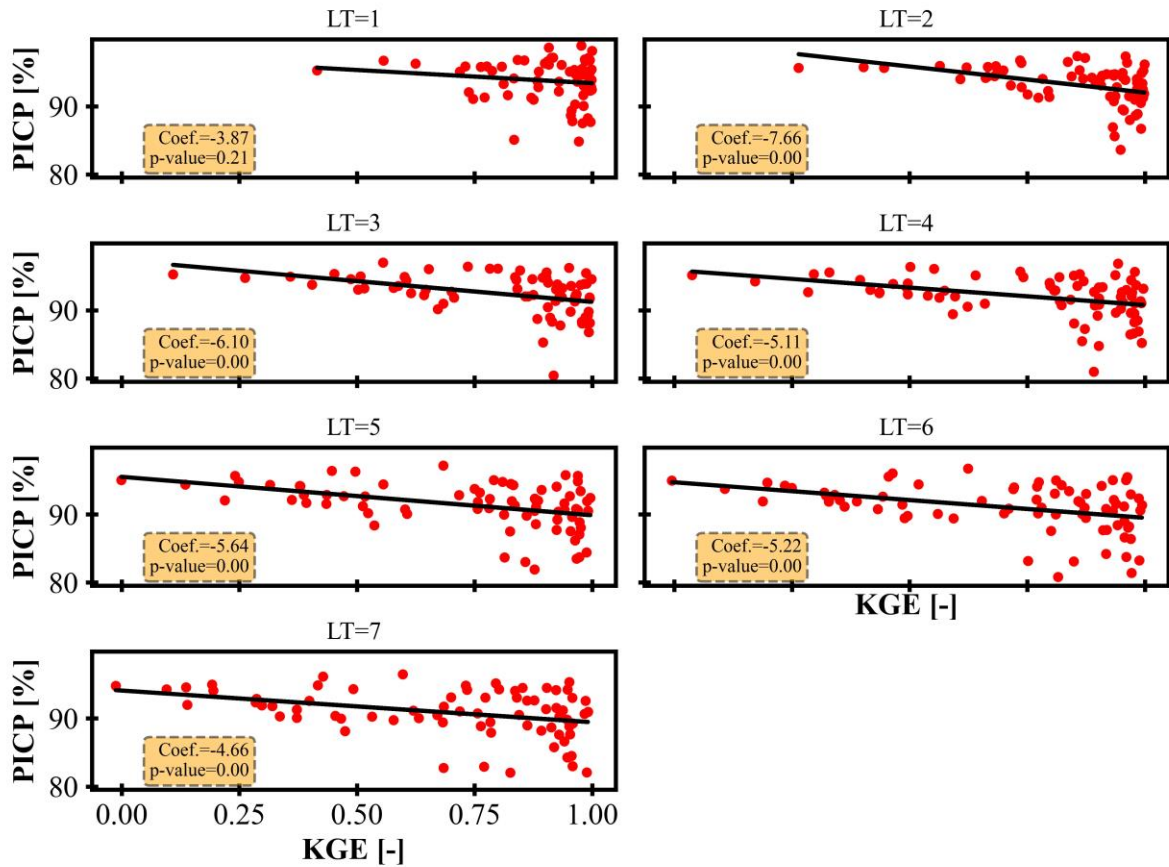


Figure 3-7. KGE vs. PICP for different LTs associated with CVAEQ. The slope coefficients and their associated p-values are shown in the box.

3.5 Conclusion, Limitations, and Future Work

This research introduced a specific generative deep learning (GDL) model, conditional variational auto-encoder (CVAE), for probabilistic multi-step ahead streamflow forecasting. The model was tested on a large sample set (75 Canadian basins) and compared to two advanced quantile-based deep learning (DL) models, one of which (QCVAE) was introduced for the first time in the hydrological forecasting literature. All models were developed using long short-term memory network (LSTMs) blocks along with a dense network layer. This study developed, evaluated, and applied over 9000 models based on different input variable sets, basins, and model structures. It was found that the quantile-based DL models demonstrated higher forecast accuracy, although CVAE produced more reliable forecasts. The performance of the forecasting models was also compared to a probabilistic streamflow simulation

model. Using the same input variable set (only meteorological data) as the simulation models, it was observed that the forecast models outperformed the simulation model in at least 20 % of the basins. While this study aimed to fill the knowledge gap in probabilistic DL-based hydrological forecasting, future research should address certain limitations.

Streamflow in most of the basins was influenced by the snowmelt process and exhibited high autocorrelation, which may limit the generalizability of the findings to other locations (e.g., semi-arid regions). Therefore, assessing the models' performance using a more diverse dataset, such as the Catchment Attributes and MEteorology for Large-sample Studies (CAMELS) dataset, is crucial to account for variability across different hydrometeorological regimes. Using large sample datasets (e.g., CAMELS), also allows for using the proposed GDL models to establish new benchmarks by forecasting other hydrological variables (e.g., precipitation).

Furthermore, analyzing the relationship between the autocorrelation of the streamflow time-series and the models' performance in each basin confirmed that selecting a large look-back (LB) period for the LSTM-based models does not necessarily improve performance. Thus, there is a need for a comprehensive investigation to determine an appropriate look-back period and select essential input variables for encoder-decoder models, which is currently missing in the literature.

It was found that the loss function played an especially significant role in the selection of optimal hyperparameters for the CVAE and QCVAE models. Thus, future research could explore in detail how to identify an optimal loss function for deep learning forecast models. Lastly, the flexible structure of the proposed GDL model (i.e., CVAE) enables various modifications, such as deriving non-parametric predictive distributions using flow-based modeling (Dumas et al., 2022), which represents an exciting topic to explore in future hydrological forecasting studies.

Chapter 4

Hierarchical Deep Learning for Consistent Multi-Timescale Hydrological Forecasting

This chapter is based on a manuscript of the same title submitted to the journal, *Water Resources Research*. Some adjustments to the manuscript have been made to ensure its better integration with the thesis.

Jahangir, M.S. and Quilty, J., 2024. Hierarchical deep learning for consistent multi-timescale hydrological forecasting. *Water Resources Research*, Under review.

Summary

Deep learning (DL) models are increasingly being used for hydrological forecasting across various timescales (hourly, daily, etc.). However, one of the main challenges with multi-timescale (MTS) DL-based hydrological forecasting is the potential inconsistency (discrepancy) between forecasts across different timescales. Inconsistent MTS forecasts can be especially problematic when decision-making relies on forecasts across different timescales. This chapter introduces a hierarchical deep learning (HDL) model incorporating temporal hierarchical reconciliation (THR) with DL models for consistent MTS streamflow forecasting. HDL is developed, deployed, and tested for multi-step (seven days ahead), MTS (daily and weekly) streamflow forecasting using over 400 catchments selected across continental United States. HDL is a generative DL model based on long short-term memory networks and normalizing flows (NFs). NFs are particularly effective for streamflow forecasting as they can estimate the probability distribution of the forecast sequence using a nested conditional distribution. HDL improved the Nash Sutcliffe efficiency (NSE) of weekly streamflow forecasts by up to 49.7% compared to base (individual) weekly forecasts and produced substantially more accurate forecasts at the weekly scale than bottom-up daily forecast aggregation. HDL performance is influenced by both the THR formulation and the accuracy of the forecasts at different timescales. Given the nascence of THR in hydrology, HDL is expected to set a benchmark for future MTS hydrological forecasting efforts.

4.1 Introduction

In water resource management, hydrological forecasts play a major role in guiding informed decision-making (Delaney et al., 2020). Accurate hydrological forecasts have substantial importance across various sectors. Civil protection agencies rely on them to anticipate and prepare for floods and droughts, safeguarding communities (Myronidis et al., 2018). Reservoir managers use these forecasts to optimize water storage and release strategies (Shang et al., 2019), ensuring a consistent water supply for agricultural activities (Liu et al., 2017), hydropower generation (Shang et al., 2019), and drinking water needs (Bakker et al., 2013). Watershed managers also depend on these forecasts to manage drought risks and maintain the health of aquatic ecosystems (Navarro-Ortega et al., 2015).

Hydrological prediction models, including simulation and forecasting models (Beven and Young, 2013), are commonly classified into process-based (physical, conceptual) and statistical models. Of the latter, the use of deep learning (DL) for hydrological prediction has gained substantial momentum in the last decade due to an increase in data availability and benchmark datasets (e.g., Caravan, Kratzert et al., 2023), open-source libraries that have simplified model development (e.g., Tensorflow, PyTorch), and affordable access to computational resources, such as graphic processing units (GPUs). The hydrology domain is beginning to explore a variety of DL models native to other fields (Devlin et al., 2019), with a recent focus on Transformers (Vaswani et al., 2017), which have shown promise for hydrological prediction (Amanambu et al., 2022; Rasiya Koya and Roy, 2024; Yin et al., 2022). Despite the promising results of more recent DL models, long short-term memory networks (LSTMs) continue to serve as the preferred DL model within the hydrology community (Kratzert et al., 2019, 2018; Nearing et al., 2024; Nevo et al., 2022; Xiang et al., 2020; Yin et al., 2021).

LSTMs excel at modeling the non-linear response of watersheds to meteorological forcing. Additionally, their ability to dynamically process information mimics the real-world response of a watershed (Alizadeh et al., 2021; Kratzert et al., 2019, 2018). The use of LSTMs within an encoder-decoder (ED) framework, where the input data (e.g., meteorological forcing) is mapped to a latent space (encoded), and then remapped to the target (e.g., streamflow) space (decoded), has shown substantial potential for hydrological prediction (Jahangir et al., 2023; Jahangir and Quilty, 2024; Kao et al., 2020; Nevo et al., 2022).

Prior research leveraging DL for hydrological prediction has predominantly concentrated on the daily timescale. However, applications in reservoir operation (Biswas et al., 2021), wetland

management (Greenberg et al., 2015), and agricultural drought monitoring (Sharma and Panu, 2014) necessitate predictions at coarser temporal scales, ranging from weekly to monthly intervals. To the authors' best knowledge, a comprehensive investigation into the efficacy of DL models like LSTMs for weekly hydrological prediction remains understudied in the literature. Consequently, this research develops and evaluates state-of-the-art DL models for the weekly timescale.

Only two studies have utilized a single DL model for multi-timescale (MTS) hydrological prediction (Gauch et al., 2021; Wilbrand et al., 2023). Both studies considered the multi-timescale LSTM (MTS-LSTM), where a single LSTM model was used to predict hourly and daily runoff simultaneously. Gauch et al. (2021), the developers of MTS-LSTM, demonstrated the efficacy of LSTMs for MTS runoff prediction, highlighting the adaptability of LSTMs for processing sub-daily data and the importance of integrating cross-timescale consistency into the loss (objective) function. In other words, the authors sought to ensure the sub-daily predictions aggregated to the daily predictions (hereafter, referred to as consistency) by including a penalty in the loss function.

However, a significant limitation of employing a single model for MTS prediction is the lack of consistency across different timescales, revealed as discrepancies where, for example, the sum of hourly predictions fails to correspond to the daily prediction. Gauch et al. (2021) proposed regularization of the loss function by penalizing these inconsistencies, leading to enhancements in prediction accuracy. While regularizing the loss function can reduce the inconsistencies of predictions across timescales, it does not reduce them to zero (as they remain unconstrained).

The inconsistency in forecasts across MTS is a well-known problem in time-series forecasting, as different models are often adopted for forecasting at different timescales (Athanasopoulos et al., 2017; Wickramasuriya et al., 2019). The resulting inconsistencies amongst forecasts at different timescales can lead to sub-optimal planning and management of scarce resources (Hyndman et al., 2011; Jahangir and Quilty, 2022). Temporal hierarchical reconciliation (THR) methods (Athanasopoulos et al., 2017) were specifically formulated to solve the latter problem. Thus, the present study builds upon the work of Gauch et al. (2021) by proposing a single DL (LSTM-based) model for MTS hydrological prediction that solves the inconsistency problem across different timescales utilizing THR methods. While the work of Gauch et al. (2021) considered runoff prediction, this study focuses on streamflow forecasting (obtaining future values). In what follows, the term forecasting will be used due to the case study adopted herein, and prediction will only be mentioned where necessary.

THR methods ensure consistency between forecasts generated by various models, each operating at distinct timescales. Notably, the approach applies to any variable that can be aggregated across different timescales (e.g., precipitation, evaporation, streamflow). Consequently, it holds significant promise for various water resource management applications, including reservoir operations, irrigation scheduling, and water supply planning (Jahangir and Quilty, 2022). This study examines the efficacy of DL-based THR for weekly streamflow forecasting using daily data. In this setting, multi-step (seven-day ahead) daily streamflow forecasts and (single-step) weekly forecasts are produced. The proposed DL-based THR model simultaneously generates streamflow forecasts for both timescales and ensures that the seven days ahead forecasts are consistent with the weekly forecasts.

Various reconciliation methods exist in the literature. Traditional reconciliation methods include the bottom-up (BU) and top-down (TD) methods (Dangerfield and Morris, 1992). The BU method sums the original (base) forecasts at lower levels of the temporal hierarchy (finer timescales) to obtain forecasts at higher temporal hierarchy levels (coarser timescales). The TD method disaggregates higher-level forecasts (at coarser timescales) to obtain lower-level forecasts (at finer timescales). An analogy between the BU and TD methods and the temporal upscaling and disaggregation methods adopted in hydrology, respectively, can be found in Wang et al. (2011) and Tsoukalas et al. (2019).

THR methods, a particular family of hierarchical reconciliation (methods), have emerged as superior approaches (compared to BU and TD) due to their incorporation of information across multiple timescales during the reconciliation process (Athanasopoulos et al., 2017; Hyndman et al., 2011; Jahangir and Quilty, 2022; Punia et al., 2020; Spiliotis et al., 2020). Previous research suggests that THR methods can improve the accuracy of base forecasts and ensure consistency across timescales. Generally, it has been shown that forecasts at higher levels of the temporal hierarchy (coarser timescales) tend to be more accurate after THR (Athanasopoulos et al., 2017). Additionally, it is understood that the time-series' characteristics (e.g., auto-correlation) can influence the effectiveness of THR (Nystrup et al., 2020).

THR methods have been investigated previously in various fields, including: energy (Nystrup et al., 2020; Spiliotis et al., 2020; Yang et al., 2017), supply chains (Punia et al., 2020), tourism (Nikolaos Kourentzes and Athanasopoulos, 2019), and water resources (Jahangir and Quilty, 2022). Additional information regarding THR methods and their variations can be found in Hyndman et al. (2011) and

Athanasopoulos et al., (2017). Some recent advancements in THR, namely, machine learning (ML)-based THR, are discussed in Section 4.2.1

4.1.1 Contribution and Outline

In the current study, autoregressive normalizing flows (NFs), a class of flow-based generative DL models (Dumas et al., 2022), are introduced and deployed for consistent MTS hydrological forecasting. The model simultaneously produces forecasts at daily and weekly scales, with a one-week (seven days) forecast horizon. Instead of using THR as a post-processing tool to make the forecasts consistent (the typical approach) or regularizing the objective function for minimizing the disparity between the timescales (Gauch et al., 2021), the reconciliation is conducted implicitly through the DL model. The result is a DL “pipeline” for consistent MTS hydrological forecasting, named hierarchical DL (HDL). In other words, a single model is developed whose parameters are optimized end-to-end by directly minimizing a loss function based on consistent forecasts. Although comparable practices have been undertaken in domains beyond water resources and hydrology (e.g., Rangapuram et al. 2021), the HDL follows an approach different than prior research (see Section 4.2 and Section 4.4 for details).

For a detailed examination of HDL, the performance of the model is assessed on more than 400 catchments across the continental United States (CONUS) using data from the Catchment Attributes and MEteorology for Large-sample Studies (CAMELS, Addor et al., 2017; Newman et al., 2015) and Caravan (Kratzert et al., 2023) datasets. The authors’ previous study (Jahangir and Quilty, 2022) was the first to explore THR in the water resources domain, where a case study on precipitation forecasting was undertaken using common statistical time-series models. No previous work has examined DL-based THR for hydrological or water resources forecasting, let alone using a large sample case study.

This study’s main contributions are as follows:

- 1- Development of the HDL model for consistent hydrological forecasting. More specifically, the exploration of how embedding THR in the DL model can be leveraged to obtain consistent forecasts across daily and weekly timescales and potentially improve the forecasts’ accuracy. While this study focuses on streamflow, without loss of generality, the methodology applies to other variables important to water resources and hydrology (e.g., precipitation and evaporation). Such an approach aims to generate consistent forecasts across multiple timescales, supporting consistent decision-making in operation, management, and planning of water resources systems.

- 2- Introduction of autoregressive NFs, for hydrological forecasting. The efficacy of NFs is assessed by benchmarking it against two LSTM variants.
- 3- Assessment of the impact of different THR methods (embedded in the HDL model) on the performance of the final reconciled forecasts. Of note, the HDL model is capable of probabilistic forecasting and, consequently, can easily be extended to non-parametric probabilistic THR (PTHR; Jeon et al., 2019; Olivares et al., 2024). The term non-parametric reflects that no specific distribution is assumed for the joint distribution of the forecasts (within each level and across the temporal hierarchy).

The current study can be classified as an ad hoc benchmarking and model intercomparison study (Klotz et al., 2022). This is because standard datasets are adopted (CAMELS and Caravan), models are developed, evaluated, and applied based on standard practices in the ML community (e.g., data splitting, standardization), and the results are disseminated to the public (see Nearing et al., 2018).

The rest of the chapter is organized as follows: Section 4.2 reviews ML-based THR and flow-based DL. Section 4.3 introduces the data used for model development, evaluation, and deployment. Section 4.4 provides details regarding the model development and evaluation criteria, followed by Section 4.5, where the results are presented and discussed. Finally, in Section 4.6, concluding remarks are given, research limitations are presented, and future work is recommended.

4.2 Related Work

This section offers a brief review of ML-based THR and flow-based generative DL models, with an emphasis on NFs. While ML-based THR has not yet been explored in hydrology or water resources, a single study (Zhou et al., 2023) has explored flow-based models in this domain, where the authors used flow-based generative DL models for reservoir inflow forecasting. The approach herein differs from Zhou et al. (2023) as this study uses a different type of NFs, namely, autoregressive NFs (see Section 4.4 for details).

4.2.1 Machine Learning-Based Reconciliation

Mancuso et al. (2021) forecasted time-series by using DL to extract features from the time-series at all hierarchy levels. Their model incorporated explanatory variables to improve the forecasting and THR process. In contrast to traditional methods that conduct THR as a post-processing step, their approach was akin to that of Gauch et al. (2021), where the loss was regularized (though consistent

forecasts were not guaranteed). In their approach, the focus was on generating disaggregated time-series from aggregated ones. In other words, the authors trained a DL model as a surrogate model for time-series disaggregation, where the aggregated time-series were inputted into the DL to produce the disaggregated time-series. Unlike their approach, which requires disaggregating the time-series one level at a time, the HDL proposed herein generates forecasts for all levels of the hierarchy simultaneously. Additionally, HDL leverages the THR, that has been shown to be superior to traditional reconciliation methods that focus solely on specific levels of the hierarchy (Hyndman et al., 2011).

Spiliotis et al. (2021) introduced a ML-based approach for THR. In their approach, a chosen ML model (e.g., neural network, random forest) was used to predict reconciled forecasts at the most disaggregated level using base forecasts from all hierarchy levels, followed by applying the BU method to obtain reconciled forecasts at all levels. Their method was tested on two datasets, confirming its superiority over existing THR methods. One advantage of their approach is the non-linear combination of forecasts (unlike most THR methods) to obtain reconciled forecasts at the most disaggregated level. The proposed model (HDL) differs from theirs as a single DL model is developed for MTS forecasting that seamlessly integrates THR to ensure forecast consistency (see Section 4.4).

A single ML/DL model for consistent MTS forecasting is not unprecedented. For instance, Rangapuram et al. (2021) introduced a neural network model that merged the forecasting and reconciliation stages into a unified framework, offering an end-to-end solution that obviated the need for a separate post-processing step. Additionally, the model was capable of probabilistic forecasting. While HDL shares similarities with Rangapuram et al. (2021), HDL can perform dynamic THR. In other words, the combination of base forecasts is time-variant in the HDL model, which is not the case in Rangapuram et al. (2021).

A limited number of studies have attempted to make THR dependent on newly available observations (i.e., dynamic THR). Charotia et al. (2022) used reinforcement learning to create a dynamic (modified version of) THR. Their approach used observations from fine timescales to correct coarser timescale forecasts. This approach differs from ‘traditional’ THR, which performs THR solely based on forecasts across all timescales. Bergsteinsson et al. (2021) introduced adaptive THR, where new observations were used to estimate forecast errors to combine base forecasts dynamically. By requiring estimates of the forecast error, an additional source of uncertainty is added when dynamically

combining base forecasts. This additional source of uncertainty can be avoided if THR is based solely on the new observations, as in HDL.

4.2.2 Flow-Based Generative Deep Learning

Generative DL represents a recent resurgence in a field established in the 1980s. This resurgence is driven by the availability of large datasets, sophisticated DL architectures, and easily accessible GPUs (Dumas et al., 2022). Generative DL employs deep neural networks to model and replicate the data generation process, aiming to generate entirely new data that closely mirrors the original dataset (Bond-Taylor et al., 2022). These models, characterized by their ED architecture, possess the flexibility to capture intricate relationships within data, which is particularly advantageous for modeling time-series in diverse domains (Cui et al., 2022; Jahangir and Quilty, 2024).

One of the generative DL models that has provided impressive results across many applications, such as speech recognition (Oord et al., 2018), load prediction (Ge et al., 2020), energy forecasting (Dumas et al., 2022), and reservoir inflow forecasting (Zhou et al., 2023), is NF. NFs represent a novel category of probabilistic generative DL models that have recently garnered growing attention within the ML community.

NFs function by generating a sequence of transformations, known as a flow, from a probability density function that is analytically known, such as a Gaussian distribution, to a more intricate target distribution (Bond-Taylor et al., 2022; Dumas et al., 2022). NFs, functioning as neural network-based probability density estimators, distinguish themselves from other generative DL models like variational autoencoders (VAEs) and generative adversarial networks (GANs) by their capacity to offer precise estimations of probability density function directly (Papamakarios et al., 2018). This attribute renders them especially suitable for applications requiring the explicit estimation of probability density functions rather than the generation of new data, as in scenario generation (e.g., synthetic data generation, data augmentation; Papamakarios et al., 2018).

Different methods exist for building NF models. The most common methods are inverse autoregressive flow (Kingma et al., 2017), non-linear independent component estimator (Nice; Dinh et al., 2015), real-valued non-volume preserving method (Real NVP; Dinh et al., 2017), and masked autoregressive flow (MAF; Papamakarios et al., 2018). This study adopts MAF, which has been shown to outperform its predecessors, such as Nice and Real NVP (Papamakarios et al., 2018). MAF uses a

masked autoencoder for distribution estimation (MADE) as a fundamental component (Germain et al., 2015). Additional details on the NF model adopted in this study are provided in Section 4.4.

4.3 Data and Case Study

In this study initially, 421 catchments were selected from the CAMELS dataset (Newman et al., 2015). The CAMELS dataset encompasses 674 basins in the CONUS with minimal human interference. These basins are categorized into 18 hydrological units (HUs) based on the U.S. Geological Survey (USGS) HU map (Addor et al., 2017). Two types of daily data were used as input to the models for forecasting weekly streamflow at two timescales: daily (seven days/steps ahead) and weekly (one step ahead).

The first dataset is Daymet. Following Yin et al. (2021) and Jahangir et al. (2023), lagged Daymet variables, including precipitation, maximum and minimum air temperature, shortwave downward radiation, vapor pressure, and streamflow observations, were used as input to the models. A fixed lag (look-back period) of 365 days was chosen, consistent with prior research that used LSTM models for hydrological prediction (e.g., Jahangir and Quilty, 2024; Kratzert et al., 2019, 2018). This involved lagging Daymet (meteorological) and streamflow variables ($\mathcal{F}_D$) by 365 days (IL_D), with the resulting sequence (number of samples $\times 365 \times 6$), fed into the DL models.

As no real-time meteorological forecast data were available for the selected catchments, European ReAnalysis (ERA5)-Land meteorological and model state reanalysis data from Caravan were used as an alternative to real-time meteorological forecasts. In total, 37 variables ($\mathcal{F}_E$) from ERA5-Land were obtained from the Caravan dataset for the selected catchments and were used as the secondary inputs (number of samples $\times 7 \times 37$) to the DL models; where a forward time lag of 7 days (IL_E) is used for the 37 ERA5-Land variables (i.e., $\mathcal{F}_E$) to be consistent with the seven days ahead streamflow forecasts. The only ERA5-Land variable excluded was minimum shortwave radiation (as it consistently equaled zero). For more details regarding the ERA5-Land variables, interested readers can refer to Kratzert et al. (2023) while Section 4.3 shows how the data described above are used in HDL.

Two criteria were employed to select the initial subset of catchments (421). The first criterion removed catchments from further analyses if they contained consecutive missing values in the streamflow record. This was done to avoid the imputation of consecutive missing records. The second

criterion removed catchments from further analyses if they did not contain at least 10,000 records for model development, deployment, and testing. To ensure consistency between the ERA5-Land dataset and the Daymet dataset, data spanning from June 1984 to December 2012 was used. The dataset was divided into three subsets: training (80%), validation (10%), and test (10%). The training and validation sets were used for network training (optimization) and hyperparameter optimization, respectively. The test set, unseen by the DL models during training and validation, was employed for performance evaluation. All results presented hereafter pertain to the test set.

Through preliminary experiments with both input datasets, it was found that using ERA5- Land data substantially decreased the performance of the models compared to using Daymet data (please see Appendix C, and Section 4.5 for details). As one of the main goals of this study is to evaluate the efficacy of DL-based THR, 19/421 catchments that showed very poor performance on the validation set were excluded. Appendix C includes the distribution of gauges associated with the final set of catchments (402) along with the number of catchments within each HU. The HUs with the highest and lowest number of catchments in the final set were HU 3 (south Atlantic Gulf) with 57 and HU 9 (Souris–Red–Rainy region) with only two, respectively.

4.4 Methodology

This section begins with an overview of THR fundamentals. Afterwards, details about HDL model development are given as well as a description of the experiments. Finally, the evaluation criteria used for model comparison is provided.

4.4.1 Temporal Hierarchical Reconciliation (THR)

In what follows, the weekly streamflow forecasting task is defined as a temporal hierarchy (Figure 4-1). The streamflow forecast values are indicated by $\hat{Q}$, where the superscript and subscript show the temporal scale and forecast horizon, respectively. For example, $\hat{Q}_5^d$ is the five-step ahead daily forecast. Generally, forecasts at each level can be produced individually (a model for each forecast horizon) or simultaneously (a single model for all forecast horizons). The latter can be referred to as a seamless forecast (McInerney et al., 2022). In this research, ED-based (sequence-to-sequence) LSTM models were used. Hence, given the sequence-to-sequence nature of the models, all the daily forecasts are obtained simultaneously. The forecasts produced at each level/timescale (before THR) are called original forecasts (ORFs) or base forecasts. The objective is to ensure that the sum of the forecasts at

both daily and weekly scales are consistent. This process (THR) guarantees that the weekly forecast matches the total of the daily forecasts (i.e., forecasted streamflow volumes for the next seven days match the weekly streamflow forecast), ensuring consistency in mass conservation.

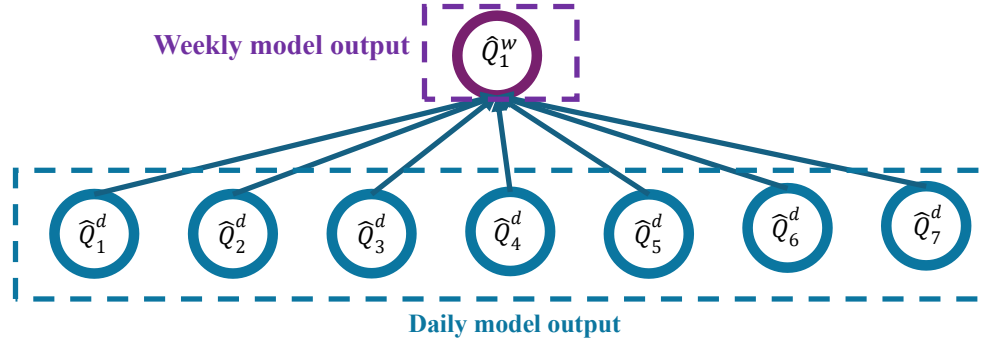


Figure 4-1. The defined hierarchy. The lower level is associated with the daily scale, while the upper level is associated with the weekly scale.

In addition to investigating the BU method for reconciliation, this research adopts three other THR methods. The first method, ordinary least squares (OLS), was proposed by Hyndman et al. (2011). The two other methods are structural scaling-based weighted least squares (WLS_S; Athanasopoulos et al., 2017), and dynamic scaling-based weighted least squares (WLS_D). The effectiveness of OLS and WLS_S for THR has already been shown by previous research (e.g., Yang et al., 2017; Jahangir and Quilty, 2022; Nystrup et al., 2020), but has yet to be incorporated in a DL model for hydrological forecasting. Furthermore, WLS_D is a dynamic DL-based THR method yet to be explored in the fields of hydrology and water resources management. While other THR methods exist (see Nystrup et al., 2020; Wickramasuriya et al., 2019 for details) and have been shown to be effective, these additional methods are out of the scope as the main focus of this research is on highlighting the benefits of incorporating THR with DL models for hydrological forecasting. The formulation of the adopted methods is provided in the following subsection.

4.4.1.1 Temporal Hierarchical Reconciliation

Consider all streamflow observations at both daily and weekly levels/timescales stacked in a single vector:

$$\mathbf{Q}_i = \mathbf{S}\mathbf{Q}_i^1 \quad \text{Eq 4-1}$$

where $\mathbf{Q}_i$ is the streamflow observations for the whole hierarchy (both levels/timescales) associated with the i -th sample, $\mathbf{S}$ is the summation matrix, and $\mathbf{Q}_i^1$ is the vector containing daily streamflow (first level) observations associated with the i -th sample. The matrix $\mathbf{S}$ is associated with the designated weekly hierarchy in Figure 4-1 and is defined as:

$$\mathbf{S} = \begin{pmatrix} \mathbf{I}_1 \otimes \mathbf{1}_7 \\ \mathbf{I}_7 \otimes \mathbf{1}_1 \end{pmatrix} \quad \text{Eq 4-2}$$

where $\mathbf{I}_d$ is the identity matrix of order d , $\mathbf{1}_j$ is a length- j vector of ones, and $\otimes$ is the Kronecker product (Werner et al., 2008).

Following the formulation provided by Athanasopoulos et al. (2017), the regression representation of forecasts based on THR is given as (Athanasopoulos et al., 2017):

$$\hat{\mathbf{Q}}_i = \mathbf{S}\boldsymbol{\beta}_i + \epsilon \quad \text{Eq 4-3}$$

where $\hat{\mathbf{Q}}_i$ is the vector containing forecasts across the entire hierarchy for the i -th sample), ϵ is the error associated with THR that has zero mean and unknown covariance matrix ($\boldsymbol{\Sigma}$), and $\boldsymbol{\beta}_i$ is the unknown conditional mean of the forecasts at the most disaggregated level (i.e., daily timescale) associated with the i -th sample (see Athanasopoulos et al., 2017 for details). If $\boldsymbol{\Sigma}$ is known, generalized least squares can be used to obtain $\boldsymbol{\beta}_i$ using (Athanasopoulos et al., 2017):

$$\boldsymbol{\beta}_i^{\text{GLS}} = (\mathbf{S}'\boldsymbol{\Sigma}^{-1}\mathbf{S})^{-1}\mathbf{S}'\boldsymbol{\Sigma}^{-1}\hat{\mathbf{Q}}_i \quad \text{Eq 4-4}$$

Consequently, the THR forecasts are obtained via:

$$\tilde{\mathbf{Q}}_i = \mathbf{S}\boldsymbol{\beta}_i^{\text{GLS}} \quad \text{Eq 4-5}$$

where $\tilde{\mathbf{Q}}$ is the vector containing the THR forecasts. Wickramasuriya et al. (2019) proved that $\boldsymbol{\Sigma}$ is not identifiable. Hence, it should be estimated. In other words, the challenge in applying THR is estimating $\boldsymbol{\Sigma}$.

In the OLS method, $\boldsymbol{\Sigma}$ is assumed to be proportional to the identity matrix ($\mathbf{I}$; Hyndman et al., 2011):

$$\boldsymbol{\beta}_i^{\text{OLS}} = (\mathbf{S}'\mathbf{S})^{-1}\mathbf{S}'\hat{\mathbf{Q}}_i \quad \text{Eq 4-6}$$

The second method, WLS_S, estimates Σ solely based on the hierarchy structure (Athanasopoulos et al., 2017). For the two-level hierarchy in Appendix A, Σ is derived as:

$$\Sigma \approx \Lambda_s = \text{diag}(7, \underbrace{1, \dots, 1}_7) \quad \text{Eq 4-7}$$

resulting in:

$$\beta_i^{\text{WLS}_S} = (\mathbf{S}'\Lambda_s^{-1}\mathbf{S})^{-1}\mathbf{S}'\Lambda_s^{-1}\hat{\mathbf{Q}}_i \quad \text{Eq 4-8}$$

The third method WLS_D, is a modification to WLS_S where Σ is estimated as follows:

$$\Sigma \approx \Lambda_D = \text{diag}(7 + w_{1,i}^2, \underbrace{1 + w_{1,i}^1, \dots, 1 + w_{7,i}^1}_7) \quad \text{Eq 4-9}$$

where $w_{h,i}^l$ is a weight associated with level l , forecast horizon (step) h , and sample i . Stacking all weights into a single vector results in $\mathbf{w}_i$, where:

$$\mathbf{w}_i = f(\hat{\mathbf{Q}}_i) \quad \text{Eq 4-10}$$

and f is a mapping function. A single-layer neural network with rectified linear unit (relu) output was selected in this research. In general, f can be any differentiable function conditioned on the ORFs and/or HDL inputs. Also, in the proposed setting, $\mathbf{w}_i$ is only a function of the ORF values, while future work can examine including input data to the models in f as well (Eq 4-10). By estimating the weights $\mathbf{w}_i$, it becomes possible to convert THR into dynamic THR as the weights are dependent on the forecasts at time i , $\hat{\mathbf{Q}}_i$. While it is possible to estimate all the elements of Λ_D directly (instead of using the weights, $\mathbf{w}_i$), this was not done for two primary reasons. Firstly, it was found that estimating Λ_D directly led to numerical instability during training of the DL models via backpropagation. Secondly, Σ should be a positive semi-definite matrix (Athanasopoulos et al., 2017). Implementing constraints to guarantee this condition was met, resulted in poor performance in preliminary experiments. However, estimating the weights in Eq 4-10 and using them in Eq 4-9 to estimate Λ_D solved both issues.

Using Eq 4-9, the conditional mean is obtained:

$$\beta_i^{\text{WLS}_D} = (\mathbf{S}'\Lambda_D^{-1}\mathbf{S})^{-1}\mathbf{S}'\Lambda_D^{-1}\hat{\mathbf{Q}}_i \quad \text{Eq 4-11}$$

In the following subsection, the method for obtaining the ORFs at both daily and weekly timescales is explained, and the HDL model for consistent MTS streamflow forecasting is introduced.

4.4.2 Normalizing Flow (NF) for Streamflow Forecasting

An NF represents the distribution of the target, $p(\mathbf{Q})$, as an invertible differentiable transformation (flow or bijector; $\mathcal{T}$) of a known density function, $p_u(\mathbf{u})$ (Dumas et al., 2022; Germain et al., 2015; Papamakarios et al., 2018).

$$\mathbf{Q} = \mathcal{T}(\mathbf{u}); \mathbf{u}, \mathbf{Q} \in \mathbb{R}^{\mathcal{M}} \quad \text{Eq 4-12}$$

In its original setting, $p_u(\mathbf{u})$ is pre-defined by the user and selected so it can be easily sampled from and evaluated for different values. Usually, a Gaussian or a uniform distribution is used for $p_u(\mathbf{u})$ (Papamakarios et al., 2018). In this research, $p_u(\mathbf{u})$ is selected as a multivariate Gaussian conditioned on the inputs; where $\mathbf{u}$ is the output of the network encoder (see details in the following subsection):

$$\mathbf{u} = f(\chi) \sim N(\boldsymbol{\mu}_u, \boldsymbol{\Sigma}_u | \chi) \quad \text{Eq 4-13}$$

where χ is the input to the DL models (past streamflow and meteorological forcing). Based on the invertibility assumption of the transformation function $\mathcal{T}$, the target distribution is expressed as (Papamakarios et al., 2018):

$$p(\mathbf{Q}) = p_u(\mathcal{T}^{-1}(\mathbf{Q})) \left| \det \left(\frac{\partial \mathcal{T}^{-1}}{\partial \mathbf{Q}} \right) \right| \quad \text{Eq 4-14}$$

where $\det \left(\frac{\partial \mathcal{T}^{-1}}{\partial \mathbf{Q}} \right)$ represents the determinant of the Jacobian matrix. Two conditions for $\mathcal{T}$ should be met so that $\mathcal{T}$ is tractable (Papamakarios et al., 2018). Firstly, $\mathcal{T}$ should be easily invertible. Secondly, the determinant of Jacobian matrix must be (efficiently) computable. In general, $\mathcal{T}$ could be defined as a composition of any number of transformation functions ($\mathcal{K}$) with the two latter properties (Papamakarios et al., 2018). Considering k compositions for $\mathcal{T}$, the target distribution can be estimated via (Papamakarios et al., 2018):

$$\log p(\mathbf{y}) = \log p_{u_0}(\mathbf{u}_0) - \sum_{ii=1}^{\mathcal{K}} \log \left(\left| \det \left(\frac{\partial \mathcal{T}_i}{\partial \mathbf{u}_{ii-1}} \right) \right| \right), \mathbf{u}_{ii} = \mathcal{T}_{ii}(\mathbf{u}_{ii-1}), ii = 1, \dots, \mathcal{K} \quad \text{Eq 4-15}$$

This study adopts MAF as the transformation function ($\mathcal{T}$). MAF has been shown to outperform its predecessors (Papamakarios et al., 2018). MAF utilizes a composition of masked autoregressive artificial neural networks (i.e., MADE) for transformation (Germain et al., 2015). Preliminary experiments demonstrated that instead of using multiple transformations, a single transformation ($\mathcal{K}=1$)

using a larger neural network (multiple layers) resulted in better performance. The target probability distribution (dimension $\mathcal{M}$; where $\mathcal{M} = 7$ for daily forecasting) is decomposed into nested conditional distributions (Germain et al., 2015; Papamakarios et al., 2018):

$$p(\mathbf{Q}) = \prod_{t=1}^{\mathcal{M}} p(Q_t | \mathbf{Q}_{<t}), \mathbf{Q}_{<t} = [Q_1, Q_2, \dots, Q_{t-1}] \quad \text{Eq 4-16}$$

where the conditional distribution can be modeled as a Gaussian distribution (Papamakarios et al., 2018):

$$p(Q_t | \mathbf{Q}_{<t}) = \mathcal{N}(\mu_{Q_t}, \sigma_{Q_t}) \quad \text{Eq 4-17}$$

and μ_{Q_t} and σ_{Q_t} are obtained using invertible masked neural networks (Germain et al., 2015; Papamakarios et al., 2018):

$$\mu_{Q_t} = f_{\mu_{Q_t}}(\mathbf{u}_{<t}), \sigma_{Q_t} = f_{\sigma_{Q_t}}(\mathbf{u}_{<t}), \mathbf{u}_{<t} = [u_1, u_2, \dots, u_{t-1}] \quad \text{Eq 4-18}$$

The nested conditional distribution estimation of the target (outlined in Eqs 15-18) is aligned with the highly autocorrelated streamflow data.

The key strength of NFs lies in their ability to model complex, non-linear relationships between past and future streamflow. Unlike traditional models that rely on parametric likelihoods (e.g., mixture of Gaussian distributions), NFs can capture intricate dependencies through the nested conditional formulation. This makes them particularly adept at forecasting streamflow patterns with high autocorrelation, leading to more accurate forecasts over multiple timesteps. The interested reader can refer to Germain et al. (2015) and Papamakarios et al. (2018) for additional details on the formulation of NFs. The proposed model for consistent MTS streamflow forecasting is introduced in the next subsection.

4.4.3 Proposed Model

The HDL model that can be used for consistent daily and weekly streamflow forecasting is illustrated in Figure 4-2. The model is comprised of three main modules: the daily module (producing seven-day ahead forecasts), the weekly module (producing one-week-ahead forecasts), and the reconciliation module (that ensures consistency of the forecasts at both daily and weekly timescales, see Section 4.4.1). Daily and weekly modules are LSTM-based ED models stacked with a MADE network (see Section 4.4.2). The LSTM-based ED submodules are similar to the proposed model by Nearing et al. (2024). The formulation of LSTM is provided in Appendix C.

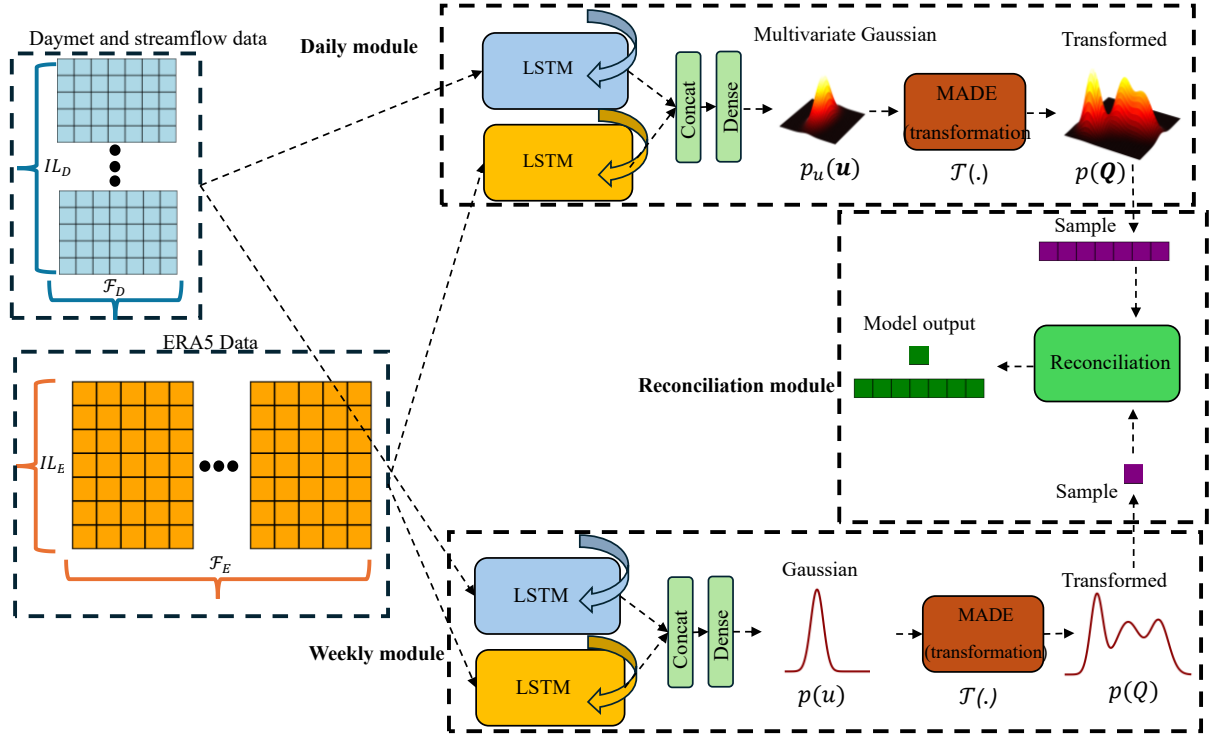


Figure 4-2. Hierarchical model for daily and weekly streamflow forecasting. $\mathcal{F}_E, \mathcal{F}_D$ indicate the number of features associated with ERA5 Land and Daymet, respectively. Similarly, IL_E and IL_D indicate the input lag (look-back period) associated with ERA5 Land and Daymet, respectively.

Transfer learning (Xu et al., 2023) was adopted to develop HDL. In particular, the daily and weekly modules were trained/optimized separately, after which their weights were frozen (fixed), and they were integrated with the reconciliation module (OLS, WLS_S , and WLS_D). An additional case was considered where the forecast modules' weights were not frozen and were then coupled with the WLS_S method. In this case, HDL underwent retraining. To better differentiate between the different variants of HDL, this approach is referred to as WLS_{SM} . In total, four different variants of HDL were explored (OLS, WLS_S , WLS_D , and WLS_{SM}), each representing a different approach to THR.

Both daily and weekly forecast modules considered hyperparameter optimization. A methodology akin to that of Jahangir et al. (2023) and Jahangir and Quilty (2024) was adopted where the hyperparameters of the models were selected using Bayesian optimization (see Wu et al., 2019). The hyperparameter optimization process included 20 iterations with each iteration optimizing the neural networks for 20 epochs, resulting in the development of 16,080 models (20 iterations $\times$ 2 timescales

$\times 402$ catchments=16080 models). The optimal hyperparameter configuration was chosen by evaluating the models' performance on the validation dataset. The model configuration associated with the highest performance (lowest root mean squared error (RMSE) across all timesteps), was fine-tuned for a maximum of 200 epochs. The list of the hyperparameters and their associated search space is indicated in Table 4-1.

Table 4-1. Hyperparameters and their associated search space (see Figure 4-2).

Hyperparameter	Type	Search space
LSTM units	Integer	[16-256]
Dense units	Integer	[8-64]
Dense activation type	Categorical	[exponential liner, linear]
MADE layer(s)	Integer	[1-3]
MADE layer size	Integer	[8-32]

While MADE was used to estimate the target probability distribution, without loss of generality, other probabilistic output (e.g., Gaussian) and deterministic (simple linear) layers can be utilized by modifying the decoder. For highlighting the efficacy of NFs for streamflow forecasting, the performance of the daily NF model (daily module) was compared to both deterministic (DT; mean estimation) and probabilistic (mixture Gaussian networks (MG)) DL benchmarks. Please refer to Klotz et al. (2022) for details regarding the MG network.

The optimization procedure was similar to Jahangir and Quilty (2024). The adaptive moment estimation (Adam) method was used as the training algorithm (see Kingma and Welling (2022)) with a RMSE loss function. A dynamic learning rate was applied for all models, initialized at $1e-3$ with a reduction factor of 0.5. This reduction was triggered every five epochs unless there was a decrease in validation loss. Additionally, to prevent overfitting, training ceased if the validation loss failed to decrease over 20 consecutive epochs. The models were developed using custom Python 3.9 scripts employing TensorFlow 2.4.1 with the Keras API. All computations were executed on the Graham cluster provided by Alliance Canada (<https://docs.alliancecan.ca/wiki/Graham>), utilizing a single GPU computational node (NVIDIA V100) equipped with 32 GB of RAM.

4.4.4 Performance Assessment

As the distribution of the target is estimated, the median of the produced forecasts was adopted for performance assessment. Common deterministic, scale-independent performance measures were adopted for evaluating the forecasts, namely, normalized root mean square (nRMSE), normalized mean absolute error (nMAE), modified Kling–Gupta efficiency (Kling et al., 2012), and Nash-Sutcliff efficiency (NSE; McCuen et al., 2006). The error metric formulae are provided in the Appendix C (see Quilty et al. (2023) for details). To identify whether the THR framework has significantly improved forecast performance compared to ORFs or not, the Diebold-Mariano (DM) test was adopted (see Diebold and Mariano, 2002).

All reported metrics are associated with the test set, data that was not used by the models during training or validation. For a fair comparison, all forecast steps for the daily model (one-seven day(s) ahead) were compared against the same target using a rolling window (Jahangir and Quilty, 2024).

4.5 Results and Discussion

This section begins by comparing the (non-reconciled) daily forecasts from the NF model (i.e., daily module output) against the deterministic (i.e., DT) and probabilistic (i.e., MG) benchmarks. Afterward, the performance of the non-reconciled weekly forecasts is assessed (i.e., weekly module output). Finally, the HDL model (see Figure 4-2) results are evaluated.

4.5.1 Non-Reconciled Daily Forecasts

The performance of NF was compared against DT and MG benchmark models. The stepwise comparison of the models based on various error metrics is depicted in Figure 4-3. The results confirm that, in general, NF outperformed DT and MG.

Considering the median of the error metrics (calculated across all timesteps and catchments), NF showed relatively similar performance to DT based on KGE. However, NF resulted in 5.36%, 4.50%, and 12.50% improvement in performance for NSE, nRMSE, and nMAE, respectively, compared to DT. A similar pattern was observed for the mean of the error metrics (across all timesteps and catchments), where NF showed 3.17%, 3.57%, 2.76%, and 4.65% improvement compared to DT for KGE, NSE, nRMSE, and nMAE, respectively (see Appendix C for details).

For all model types, a substantial performance drop was observed when the lead time was increased, especially when moving from forecast step one to two (days). For example, the NF model had a median

KGE (across all catchments) of 0.75 at step one that dropped to 0.67 for step two. The performance drop can be attributed to the decreasing effect of streamflow observations as the input to the models, where the drop in performance was related to the autocorrelation of the streamflow time-series. In other words, streamflow time-series that had higher autocorrelation experienced a lower decrease in performance as the lead time was increased. Given that ERA5 Land reanalysis data was used as a substitute for actual meteorological forecasts (equivalent to near-perfect forecasts), similar performance was observed for steps three-seven.

The forecast performance was also benchmarked against simulation (same-day prediction); where two different LSTM models (MG networks) using Daymet and ERA5 Land data were adopted for same-day streamflow prediction (please see the Appendix C for more details). In the simulation setup, lagged meteorological data (365 days) were used as model inputs, but past streamflow observations were not. The KGE comparison between two model types is presented in the Appendix C. Results confirm that using ERA 5 Land data resulted in less accurate streamflow predictions (median KGE of 0.64 compared to 0.73 when using Daymet instead), which is aligned with the findings of Clerc-Schwarzenbach et al. (2024). The comparison of the daily forecast results (Figure 4-3) with benchmark simulation models indicates that as the lead time increased, it approached that of the ERA5 Land simulation model, signifying the diminishing impact of streamflow observations on the forecasts.

4.5.2 Non-Reconciled Weekly Results

The median and mean of the error metrics associated with the weekly scale, as obtained using the weekly NF forecast module (see Figure 4-2), are presented in Appendix C. The median and mean KGE of the weekly forecasts were 0.75 and 0.65, respectively. Comparing the performance of the weekly and daily forecasts (the latter with a median and mean KGE of 0.66 and 0.65, respectively) indicates that while the weekly forecasts in general showed better performance, they also included a higher number of catchments with poor forecast accuracy similar to the mean of the daily forecasts.

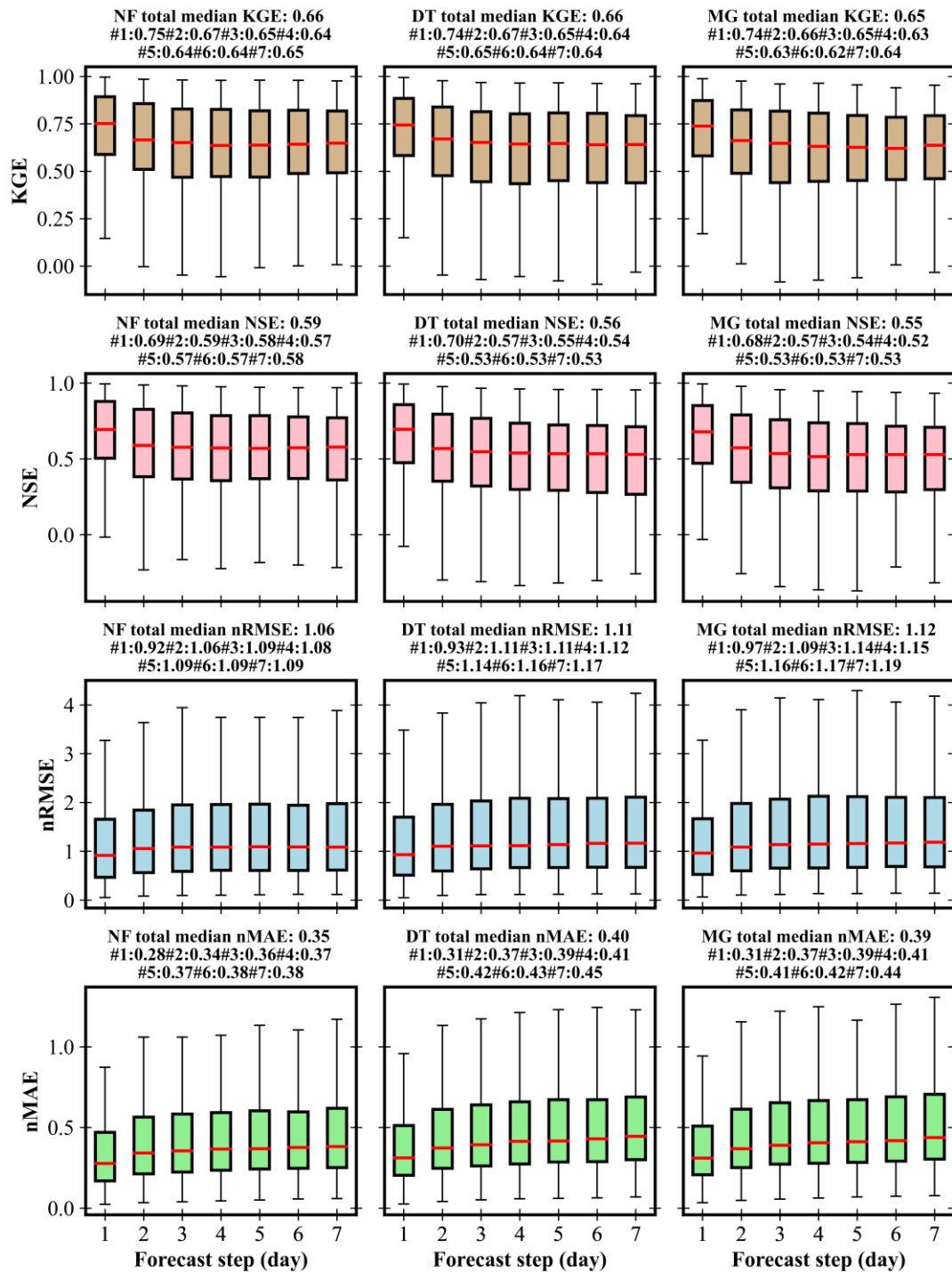


Figure 4-3. Stepwise comparison of the models' performance associated with the daily scale. The median reported at the top of each chart was calculated across all catchments and forecast steps.

As noted in the introduction section, one of the main goals of this research was to propose HDL to solve the inconsistency problem of MTS hydrological forecasting. To highlight the problem, the absolute and relative discrepancies between aggregated daily forecasts and weekly forecasts are shown in Figure 4-4. The relative discrepancy between the two timescales is derived by scaling the absolute difference by the weekly forecast value. The obtained relative discrepancies' distribution is positively skewed, with a median of 18.12% and mean of 35.26%. The relation between relative discrepancy and weekly forecast accuracy (KGE) was checked, and it was observed that a significant negative correlation exists between weekly KGE scores and relative discrepancy (at the 0.95 significance level). That is, the less accurate the weekly model, the higher the discrepancy. The HDL model can reduce both the absolute (and relative) discrepancy to zero through integrating THR seamlessly into a single DL model. In the next subsection, the results of HDL are provided.

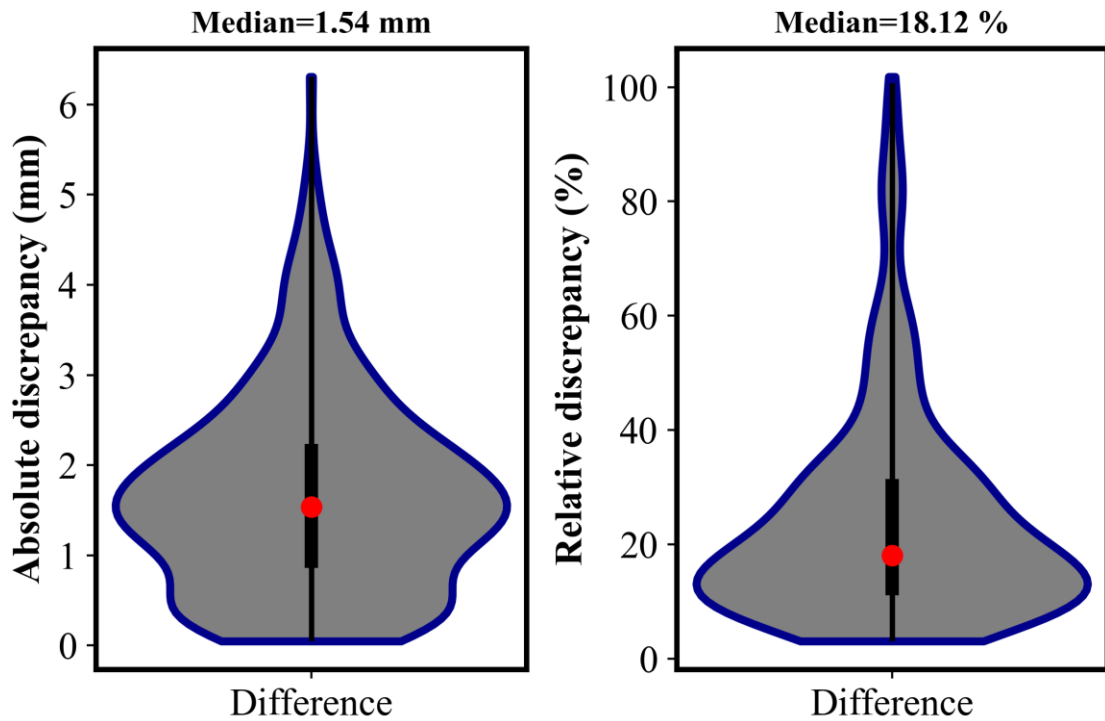


Figure 4-4. Absolute and relative discrepancy between the two timescales.

4.5.3 Reconciled Forecasts using Hierarchical Deep Learning (HDL)

For evaluating the efficacy of the proposed HDL model (that seamlessly integrates THR within a single DL model), the performance of the reconciled forecasts generated by HDL were compared against the weekly ORFs. The relative difference between the error metrics associated with HDL and ORFs is depicted in Figure 4-5. The figure shows the relative difference in error metrics for different reconciliation methods (OLS, WLS_S, WLS_D, WLS_{SM}) across different thresholds, using the BU method as another benchmark. The thresholds are based on the KGE of the ORFs at the weekly timescale across all catchments where threshold one includes all catchments whose KGE fall within the third and fourth decile, threshold two is associated with catchments whose KGE fall within the fourth and fifth decile, and so forth.

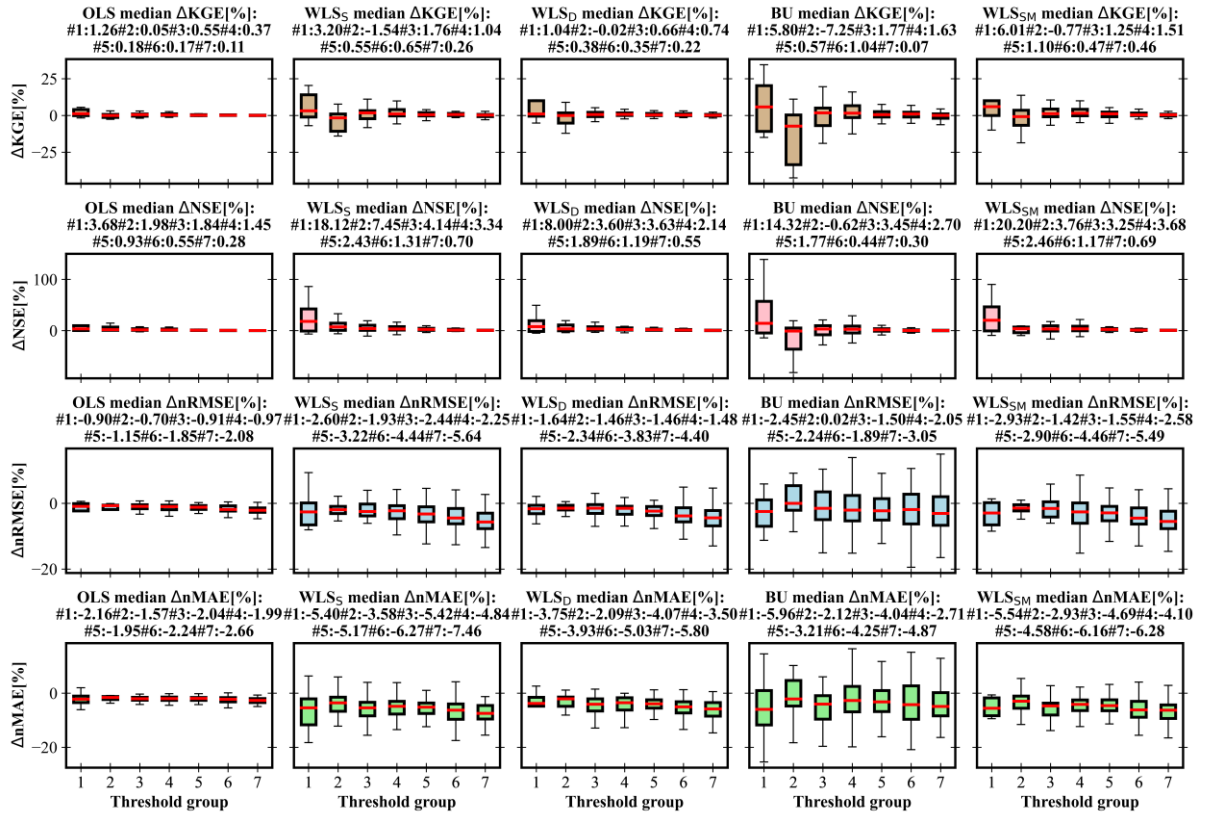


Figure 4-5. Relative difference between reconciled and original forecasts at the weekly scale. The x-axis represents the threshold group (1-7), while the y-axis represents the relative change in different error metrics compared to the ORF.

The results indicate that using HDL generally leads to better performance at the higher hierarchy level (weekly timescale), which is aligned with the findings of Jahangir and Quilty (2022) and

Athanasopoulos et al. (2017). In other words, the results confirm that for weekly streamflow forecasting, leveraging daily forecasts and reconciling them with weekly forecasts proved to be a more effective approach than solely forecasting a week ahead. Even when a non-hierarchical method, such as BU, was adopted for reconciliation, improvements in performance were noticed. It is important to note that actual meteorological forecasts were not used in the model but rather ERA5 Land data at future dates. Thus, it is expected that the performance of the different models would be lower if actual meteorological forecasts were used. However, it is not known how the gap in performance between the ORFs and the HDL would change if actual meteorological forecasts were used in the models.

The results confirm that the type of THR method substantially affects the accuracy of HDL. In general, the WLS_S method outperformed other THR methods. Conversely, employing the OLS method for THR yielded negligible improvement in forecasts. In general, except for the first threshold group (30%-40%), WLS_S, WLS_D, and WLS_{SM} demonstrated superior performance compared to the BU method. The impact of THR on the ORFs decreased as the ORFs' KGE and NSE increased. This was expected due to the negative correlation between the discrepancy between forecasts at the daily and weekly timescales (see Figure 4-4) and the KGE of the ORFs. Also, it is observed that THR resulted in no substantial improvement in KGE for the 10th decile threshold group.

For a robust comparison of the THR methods with the BU method, the significance of forecast improvement due to the THR methods compared to BU was checked using the DM test for daily and weekly timescales. A significance level of 95% was considered for the test ($\alpha = 0.05$). The results are presented in Table 4-2 and Table 4-3 for daily and weekly scales, respectively (Note: the BU method is equivalent to the ORFs at the daily timescale).

It is understood that THR had a more significant impact on the weekly timescale. For the daily timescale, except for lead times two and three, in at least 50% of catchments, no significant increase or decrease in the accuracy of the forecasts was observed (i.e., ORF and reconciled forecasts were not significantly different). This negative impact of THR, specifically on lead time two, is linked to the combined impact of THR formulation (e.g., considering similar error variance for all lead times at the same timescale) and the ORFs' error. During the reconciliation process, lead time two is more influenced by lead times one and three, where the assumed uniform error variance across lead times does not align with the substantial error increase from lead time one to two. These findings highlight

that both the THR method and ORFs' error properties (not only the amount, but the distribution across the levels/timescales) can substantially impact the final THR results.

Table 4-2. DM test percentage (%) of significant change in forecast accuracy at the daily level/timescale across 402 catchments compared to BU (ORF). The values in bold show higher percentages.

Lead time (days)	OLS		WLS _S		WLS _D		WLS _{SM}	
	Better	Worse	Better	Worse	Better	Worse	Better	Worse
1	7.71	33.33	20.65	11.19	12.44	23.38	22.14	13.43
2	5.22	84.08	12.94	56.47	8.46	68.66	10.20	65.42
3	5.72	58.71	15.92	22.89	11.69	40.05	15.42	32.59
4	6.97	43.03	23.63	15.17	12.94	29.35	20.40	21.14
5	8.96	34.58	26.37	11.19	13.68	23.13	22.64	16.42
6	8.96	33.33	27.36	10.70	12.94	21.64	23.63	13.93
7	9.20	30.35	26.37	9.45	13.18	24.13	20.65	15.67

Table 4-3. DM test percentage (%) of significant change in forecast accuracy at the weekly level/timescale across 402 catchments compared to BU. The values in bold show higher probabilities

Lead time (weekly)	OLS		WLS _S		WLS _D		WLS _{SM}	
	Better	Worse	Better	Worse	Better	Worse	Better	Worse
1	24.63	35.32	49.75	11.69	37.06	15.67	41.04	15.92

4.6 Conclusions, Limitations, and Future Work

In this research, a hierarchical deep learning (HDL) model was introduced for incorporating temporal hierarchical reconciliation (THR) with DL models (through a differentiable layer) for consistent MTS (daily and weekly) streamflow forecasting. The proposed model was developed, deployed, and tested using more than 400 catchments in CONUS selected from the CAMELS dataset, where more than 16, 000 DL models were developed and examined. Three THR methods, namely, WLS_S (structural scaling-based weighted least squares), WLS_D (dynamic scaling-based weighted least squares), and OLS (ordinary least square), were implemented in HDL and compared to the BU (bottom-up) method.

LSTM-based normalizing flows (NFs) were introduced for multi-step streamflow forecasting and were benchmarked against two other LSTM-based forecast models. NFs and the benchmark LSTM models were used to generate original forecasts (ORFs) that were used to demonstrate the utility of NFs

and evaluate the effectiveness of HDL for streamflow forecasting. The forecast horizon was set as one week, and forecasts were produced at two timescales (levels), namely, daily (seven step-ahead), and weekly (one step ahead). It was observed that a negative correlation exists between the two discrepancies between the forecasts at both timescales and the weekly ORFs' accuracy. The results confirmed that NFs show superior performance to other LSTM alternatives, highlighting the model's potential as a powerful generative DL model for hydrological forecasting. A promising avenue of research in hydrology and water resources is exploring its other variants that could lead to further improvements in performance (e.g., Real NVP (Dinh et al., 2017)).

Results confirmed the efficacy of THR methods for improving weekly streamflow forecasts. For the HDL models, WLS₅ resulted in the highest improvement in performance compared to ORFs, while deploying OLS showed the minimal improvement. While, in general, HDL performed similarly to the ORFs at the daily timescale, HDL resulted in substantially better weekly forecasts than ORFs and using the BU method. THR was impacted by both the THR method and ORF error properties at different levels. Furthermore, results showed the superiority of static THR based on the hierarchy structure compared to the dynamic THR for MTS streamflow forecasting. It is noted that the selected timescales (daily and weekly) and the states (forecast values) used for dynamic THR have an impact on the latter finding.

This research investigated MTS streamflow forecasting from a novel perspective. However, certain limitations could be addressed in future research. For example, it has been recently shown that training a single DL model on many catchments is more beneficial for hydrological prediction than catchment-wise DL model development (Nearing et al., 2024). Future research could thoroughly investigate such an approach for MTS forecasting. Furthermore, as the proposed THR layer is implemented as a differentiable layer inside the DL model, future work could integrate it with physics-based differentiable layers, such as the routing layer proposed by Bindas et al. (2024), for MTS distributed hydrological forecasting. While streamflow was used as a case study in this work, without loss of generality, a similar HDL model can be developed and deployed for other hydrological and water resources forecasting tasks, such as consistent MTS reservoir inflow forecasting that can be adopted for forecast-informed reservoir operation (Delaney et al., 2020).

Chapter 5

Conclusions, Limitations, and Future Work

Accurate and consistent hydrological and water resources forecasts at various timescales (e.g., daily, weekly) are essential for effective water resources management. Data-driven, especially DL-based, hydrological models have proven highly effective in capturing complex non-linear relationships within hydrological data, leading to more accurate simulations and forecasts compared to process-based models. This capability will continue to play a significant role in the future by assisting in sustainable water resource management, reliable hydropower generation, and mitigating the impacts of hydrological extremes.

When forecast models are developed independently for different timescales, there is no guarantee that consistent forecasts will be obtained. Significant efforts over the last two decades in the time-series forecasting community have addressed this issue, resulting in the development of THR methods. This thesis provides new insights on how THR can be incorporated with DL models for accurate and consistent MTS hydrological and water resources forecasting.

In Chapter 2, the potential of THR to produce accurate and consistent hydrological forecasts was verified through a large-scale precipitation forecasting experiment involving 84 Canadian catchments. Next, in Chapter 3, two novel state-of-the-art sequence-to-sequence PDL models were introduced and tested on more than 75 catchments across the CONUS and Canada, marking the first use of a generative DL model for probabilistic streamflow forecasting. These models demonstrated superior performance in both forecast accuracy and reliability compared to recent benchmarks. Finally, a hierarchical DL model (i.e., HDL) that integrates THR directly into the DL model architecture is introduced in Chapter 4, where its efficacy for consistent and accurate MTS streamflow forecasting was demonstrated. A major benefit of HDL is that it overcomes the need for post-processing (a requirement of almost all THR methods) by integrating THR directly in the DL model through a novel (custom) neural network layer.

The power of HDL compared to other DL benchmarks lies in its ability to leverage temporal information during training, perform probabilistic THR, and ensure forecast consistency across multiple timescales. HDL was applied for weekly streamflow forecasting (seven days ahead and one

week ahead) on more than 400 catchments across the CONUS and benchmarked against various recent DL models, highlighting its superior accuracy.

The methods proposed in this thesis are general. They can be applied to various types of hydrometeorological variables (precipitation, evapotranspiration, soil moisture, etc.), and can be coupled with different DL architectures (LSTMs, Transformers, convolutional networks, etc.). The next two sections provide concluding remarks and recommendations for future work based on the findings in each chapter.

5.1 Conclusions

Chapter 2 investigated the application of THR methods for obtaining consistent precipitation forecasts and improving the accuracy of base forecasts (ORFs) through a large-scale experiment in Canada. Three THR methods, namely, WLS_S , WLS_V , and OLS along with the simple BU method, were evaluated for reconciling the ORFs. The forecasting methods ETS, ANN, and SARIMA were selected to produce the ORFs. The forecast horizon was set to one year, with forecasts generated using a rolling window at monthly, bi-monthly, quarterly, four-monthly, bi-annual, and annual scales. The study utilized precipitation forecasts for selected catchments from the CANOPEX dataset. Specific conclusions from this chapter are outlined as follows:

- Forecast accuracy varied depending on the catchments, timescale, and forecast model. In general, ETS forecasts were more accurate than ANN and SARIMA forecasts.
- In most cases, the BU method performed the worst among the reconciliation methods.
- The WLS_S and OLS methods slightly outperformed the WLS_V method, likely due to the difference between training error and test set error. Catchments with high differences between training and testing errors, indicating model overfitting, showed decreased forecast accuracy when using the WLS_V method.
- Although THR could generally improve both accurate and inaccurate forecasts, the ORF accuracy had a much greater impact on the final reconciled forecasts than the specific THR method. Different timescales, with varying levels of accuracy, exhibited different sensitivities to THR.

Chapter 3 investigated the application of a GDL model, the CVAE, for probabilistic multi-step (seven days) ahead streamflow forecasting. This chapter addressed the gap in knowledge regarding utilization

of PDL for hydrological forecasting by introducing promising methods to the field. A large dataset encompassing 75 Canadian catchments was used to evaluate the performance of CVAE compared to two advanced quantile-based DL models. One of these quantile-based DL models, QCVAE, is a novel introduction to the field of hydrological and water resources forecasting. All models were built using LSTMs. Chapter 3 involved the development, evaluation, and application of over 9000 models, considering variations in input variable sets, catchments, and model structures. Furthermore, the findings promote the ED framework as a promising approach for probabilistic hydrological forecasting due to its ability to generate accurate and reliable forecasts. The findings emphasize the critical role of hyperparameter optimization (tuning) in achieving optimal performance from DL models when applied to hydrological forecasting. These insights can guide future research on developing even more advanced and effective PDL models.

The following conclusions are drawn from Chapter 3:

- A comparison of CVAE with two advanced quantile-based DL models revealed a trade-off between forecast accuracy and reliability. While the quantile-based models, including the novel QCVAE, achieved statistically significant improvements in point forecast accuracy metrics, CVAE demonstrated superior forecast reliability, as measured by metrics like KGE and PICP. This suggests that CVAE's forecasts better resembled the actual observed probability distribution of streamflow.
- The forecasting models, including both CVAE and the quantile-based models, substantially outperformed a traditional probabilistic streamflow simulation model, similar to the one proposed by Klotz et al. (2022). This finding highlights the effectiveness of DL models in capturing complex relationships within hydrological data for streamflow forecasting, even when limited to using only meteorological data as input.

The findings from Chapter 3 provide valuable insights into the effectiveness of GDL models for hydrological forecasting. While quantile-based models may offer slightly more accurate point forecasts, CVAE's strength lies in its ability to generate reliable probabilistic forecasts, which is crucial for risk assessment and water resource management decisions. Notably, CVAE's performance is particularly advantageous in catchments where traditional deterministic DL forecasting methods struggle. Additionally, this chapter's exploration of many models with varying configurations paves the way for

future research to optimize CVAE for specific forecasting applications (e.g., short- and long-range reservoir inflow forecasting).

Chapter 4 introduced and explored a novel HDL model that integrates THR directly within the DL model architecture. A key innovation is the custom-designed THR layer, which can be seamlessly incorporated into any DL model, like a LSTM layer or a linear layer. This integrated approach offers several advantages: it allows any model to leverage temporal information across MTS during training, perform probabilistic THR, and maintain efficiency for real-time applications. Furthermore, the HDL model utilizes autoregressive NFs, a state-of-the-art GDL model surpassing the flexibility of CVAE by enabling non-parametric estimation of the target (streamflow) probability distribution. The proposed method's efficacy was tested on more than 400 catchments across CONUS, where more than 16,000 DL models were developed, tested, and deployed. The key conclusions of Chapter 4 are as follows:

- NFs outperform other LSTM-based forecast models, demonstrating their potential as powerful GDL models for hydrological forecasting.
- HDL significantly improves weekly streamflow forecasts compared to ORFs and the BU method.
- Static THR based on the hierarchical structure proves more advantageous than dynamic THR for MTS streamflow forecasting.
- HDL generates forecasts with high accuracy, that maintain consistency across daily and weekly timescales, eliminating the need for post-processing.
- The generality of HDL makes it suitable to forecast other hydrological and water resources variables, as well as testing the approach in other disciplines like renewable energy.

By establishing a new benchmark for THR methods and offering a generalizable framework, the research presented in Chapter 4 holds substantial promise for improving decision-making in water resources management and advancing MTS forecasting across various domains.

5.2 Limitations and Future Work

While this thesis made progress in addressing several gaps in the literature related to DL-based MTS hydrological and water resources forecasting, there are shortcomings that future work should tackle. Future research should investigate several key areas to enhance hydrological and water resources forecasting accuracy and reliability. Firstly, it should explore using the best possible forecast model for each timescale, rather than a single model type (such as ETS in Chapter 2) for all timescales. Once an optimal model has been identified for each timescale, THR should be applied to the optimal models. Additionally, simultaneous spatial and temporal reconciliation (cross-temporal reconciliation) could potentially improve MTS spatio-temporal forecasting accuracy. Developing and testing new stochastic water resources forecasting frameworks based on ML, coupled with probabilistic THR methods, presents another promising avenue. This includes integrating frameworks like those proposed by Quilty and Adamowski (2020) with methods from Athanasopoulos et al. (2020). The choice of the best THR method remains case- and timescale-dependent, necessitating a thorough investigation for various case studies and THR methods to determine the most suitable methods for forecasting key hydrological and water resources variables such as streamflow, evaporation, and urban water demand.

While CVAEs (Chapter 3) and NFs (Chapter 4) have proven effective for multi-step ahead streamflow forecasting, incorporating a wider range of GDL models could be valuable. Generative Adversarial Networks (GANs) offer a promising alternative. GANs, by using adversarial training, can lead to highly realistic forecasts, potentially improving upon the performance of the chosen models in this thesis (Bond-Taylor et al., 2022). However, GANs have been proved difficult to train (Bao et al., 2022). Thus, careful attention must be given to training such models for hydrological and water resources forecasting. Another avenue to explore is diffusion models. These models learn to progressively denoise a corrupted version of the target variable(s), and have exhibited effectiveness in different time-series forecasting tasks (Rasul et al., 2021). The diffusion model is effectively "denoising" the target variable(s) to estimate their distribution. This training procedure enables the model to "understand" the underlying structure of the target, thereby equipping the model with the capability to generate new and realistic predictions. By incorporating these techniques alongside the existing models, a more comprehensive examination of GDL models for hydrological and water resources forecasting can be achieved. Such an endeavor will be useful for identifying superior forecasting techniques for the chosen forecasting task.

One limitation of this work was the catchment-wise DL model development and training strategy. Recent research suggests that training models on data from multiple catchments can lead to improved generalizability and capture broader hydrological patterns (Nai et al., 2024). While this limitation is acknowledged, it is important to note that the approach for DL model development (see Chapter 3, for example) adopted herein utilized robust hyperparameter optimization techniques. This optimization resulted in performance metrics that were similar to what might be achieved through a regional model development strategy. In unpublished experiments (tangential to this thesis), it was found that the approach adopted herein (catchment-wise development and training strategy) led to superior performance than regional models. However, it is worth recognizing that multi-catchment DL development for large models often requires very complex and expensive computational resources that might not be accessible to everyone. In this regard, the focus on catchment-wise training with efficient model architectures aimed to provide a general and accessible approach for MTS hydrological forecasting. Catchment-specific versus regional training for hydrological forecasting through DL models remains an ongoing area of research, and further investigation is required.

This thesis did not investigate using actual ensemble meteorological forecasts as input to the DL forecast models. This represents a gap in the current literature on complex DL model development for MTS forecasting. Future work can substantially benefit from exploring how DL models can leverage actual ensemble meteorological forecasts. This integration has the potential to enhance the generalizability of such models to a wide class of hydrological and water resources applications (e.g., forecast-informed reservoir operations, early flood warning systems). Furthermore, it is important to consider the combined uncertainty associated with ensemble meteorological forcings and the uncertainty quantified through the DL models (see Chapter 3). Addressing this combined uncertainty remains an active area of research, and future work should explore appropriate techniques to quantify, and reduce these uncertainties.

Addressing these areas is expected to substantially advance data-driven, especially DL-based hydrological and water resources forecasting, providing more accurate, reliable, and consistent forecasts, crucial for effective water resources management and decision-making.

References

- Abhishek, K., Kumar, A., Ranjan, R., Kumar, S., 2012. A rainfall prediction model using artificial neural network, in: IEEE Control and System Graduate Research Colloquium. IEEE, pp. 82–87.
- Abraham, B., 1982. Temporal aggregation and time series. *International Statistical Review/Revue Internationale de Statistique* 50, 285–291. <https://doi.org/10.2307/1402498>
- Addor, N., Newman, A.J., Mizukami, N., Clark, M.P., 2017. The CAMELS data set: catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences* 21, 5293–5313. <https://doi.org/10.5194/hess-21-5293-2017>
- Akaike, H., 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19, 716–723. https://doi.org/10.1007/978-1-4612-1694-0_16
- Alizadeh, B., Ghaderi Bafti, A., Kamangir, H., Zhang, Y., Wright, D.B., Franz, K.J., 2021. A novel attention-based LSTM cell post-processor coupled with bayesian optimization for streamflow prediction. *Journal of Hydrology* 601, 126526. <https://doi.org/10.1016/j.jhydrol.2021.126526>
- Amanambu, A.C., Mossa, J., Chen, Y.-H., 2022. Hydrological Drought Forecasting Using a Deep Transformer Model. *Water* 14, 3611. <https://doi.org/10.3390/w14223611>
- Apaydin, H., Sibtain, M., 2021. A multivariate streamflow forecasting model by integrating improved complete ensemble empirical mode decomposition with additive noise, sample entropy, Gini index and sequence-to-sequence approaches. *Journal of Hydrology* 603, 126831. <https://doi.org/10.1016/j.jhydrol.2021.126831>
- Arsenault, R., Bazile, R., Ouellet Dallaire, C., Brissette, F., 2016. CANOPEX: A Canadian hydrometeorological watershed database. *Hydrological Processes* 30, 2734–2736. <https://doi.org/10.1002/hyp.10880>
- Assimakopoulos, V., Nikolopoulos, K., 2000. The theta model: A decomposition approach to forecasting. *International Journal of Forecasting* 16, 521–530. [https://doi.org/10.1016/s0169-2070\(00\)00066-2](https://doi.org/10.1016/s0169-2070(00)00066-2)
- Athanasopoulos, G., Ahmed, R.A., Hyndman, R.J., 2009. Hierarchical forecasts for Australian domestic tourism. *International Journal of Forecasting* 25, 146–166. <https://doi.org/10.1016/j.ijforecast.2008.07.004>

Athanasopoulos, G., Gamakumara, P., Panagiotelis, A., Hyndman, R.J., Affan, M., 2020. Hierarchical forecasting, in: *Macroeconomic ForeCasting in the Era of Big Data*. Springer, pp. 689–719.

Athanasopoulos, G., Hyndman, R.J., Kourentzes, N., Petropoulos, F., 2017. Forecasting with temporal hierarchies. *European Journal of Operational Research* 262, 60–74. <https://doi.org/10.1016/j.ejor.2017.02.046>

Athanasopoulos, G., Hyndman, R.J., Song, H., Wu, D.C., 2011. The tourism forecasting competition. *International Journal of Forecasting* 27, 822–844. <https://doi.org/10.1016/j.ijforecast.2010.04.009>

Bakker, M., Vreeburg, J.H.G., van Schagen, K.M., Rietveld, L.C., 2013. A fully adaptive forecasting model for short-term drinking water demand. *Environmental Modelling & Software* 48, 141–151. <https://doi.org/10.1016/j.envsoft.2013.06.012>

Banihabib, M.E., Mousavi-Mirkalaei, P., 2019. Extended linear and nonlinear auto-regressive models for forecasting the urban water consumption of a fast-growing city in an arid region. *Sustainable Cities and Society* 48, 101585. <https://doi.org/10.1016/j.scs.2019.101585>

Bao, J., Li, L., Davis, A., 2022. Variational Autoencoder or Generative Adversarial Networks? A Comparison of Two Deep Learning Methods for Flow and Transport Data Assimilation. *Math Geosci* 54, 1017–1042. <https://doi.org/10.1007/s11004-022-10003-3>

Ben Taieb, S., Bontempi, G., Atiya, A.F., Sorjamaa, A., 2012. A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *Expert Systems with Applications* 39, 7067–7083. <https://doi.org/10.1016/j.>

Bergsteinsson, H.G., Møller, J.K., Nystrup, P., Pálsson, Ó.P., Guericke, D., Madsen, H., 2021. Heat load forecasting using adaptive temporal hierarchies. *Applied Energy* 292, 116872. <https://doi.org/10.1016/j.apenergy.2021.116872>

Betts, A.K., Viterbo, P., Wood, E., 1998. Surface energy and water balance for the Arkansas–Red River basin from the ECMWF reanalysis.

Beven, K., Young, P., 2013. A guide to good practice in modeling semantics for authors and referees. *Water Resources Research* 49, 5092–5098. <https://doi.org/10.1002/wrcr.20393>

Bhatt, C.M., Rao, G.S., Diwakar, P.G., Dadhwal, V.K., 2017. Development of flood inundation extent libraries over a range of potential flood levels: a practical framework for quick flood response. *Geomatics, Natural Hazards and Risk* 8, 384–401. <https://doi.org/10.1080/19475705.2016.1220025>

Bian, C., He, H., Yang, S., Huang, T., 2020. State-of-charge sequence estimation of lithium-ion battery based on bidirectional long short-term memory encoder-decoder architecture. *Journal of Power Sources* 449, 227558. <https://doi.org/10.1016/j.jpowsour.2019.227558>

Bindas, T., Tsai, W.-P., Liu, J., Rahmani, F., Feng, D., Bian, Y., Lawson, K., Shen, C., 2024. Improving River Routing Using a Differentiable Muskingum-Cunge Model and Physics-Informed Machine Learning. *Water Resources Research* 60, e2023WR035337. <https://doi.org/10.1029/2023WR035337>

Biswas, N.K., Hossain, F., Bonnema, M., Lee, H., Chishtie, F., 2021. Towards a global Reservoir Assessment Tool for predicting hydrologic impacts and operating patterns of existing and planned reservoirs. *Environmental Modelling & Software* 140, 105043. <https://doi.org/10.1016/j.envsoft.2021.105043>

Bittelli, M., Tomei, F., Pistocchi, A., Flury, M., Boll, J., Brooks, E.S., Antolini, G., 2010. Development and testing of a physically based, three-dimensional model of surface and subsurface hydrology. *Adv. Water Resour* 33, 106–122.

Blei, D.M., Kucukelbir, A., McAuliffe, J.D., 2017. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association* 112, 859–877. <https://doi.org/10.1080/01621459.2017.1285773>

Bond-Taylor, S., Leach, A., Long, Y., Willcocks, C.G., 2022. Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 7327–7347. <https://doi.org/10.1109/TPAMI.2021.3116668>

Bozdogan, H., 1987. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika* 52, 345–370. <https://doi.org/10.1007/bf02294361>

Brochu, E., Cora, V.M., de Freitas, N., 2010. A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning. <https://doi.org/10.48550/arXiv.1012.2599>

- Brockwell, P.J., Davis, R.A. (Eds.), 2016. Introduction to time series and forecasting. Springer.
- Bukovsky, M.S., Karoly, D.J., 2007. A brief evaluation of precipitation from the North American Regional Reanalysis. *Journal of Hydrometeorology* 8, 837–846. <https://doi.org/10.1175/jhm595.1>
- Charotia, H., Garg, A., Dhama, G., Maheshwari, N., 2022. Dynamic Temporal Reconciliation by Reinforcement learning. <https://doi.org/10.48550/arXiv.2201.11964>
- Chatfield, C., 1988. What is the ‘best’ method of forecasting? *Journal of Applied Statistics* 15, 19–38. <https://doi.org/10.1080/02664768800000003>
- 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>
- Cheng, M., Fang, F., Navon, I.M., Zheng, J., Tang, X., Zhu, J., Pain, C., 2022. Spatio-Temporal Hourly and Daily Ozone Forecasting in China Using a Hybrid Machine Learning Model: Autoencoder and Generative Adversarial Networks. *Journal of Advances in Modeling Earth Systems* 14, e2021MS002806. <https://doi.org/10.1029/2021MS002806>
- Choubin, B., Zehtabian, G., Azareh, A., Rafiei-Sardooi, E., Sajedi-Hosseini, F., Kişi, Ö., 2018. Precipitation forecasting using classification and regression trees (CART) model: A comparative study of different approaches. *Environmental Earth Sciences* 77, 1–13.
- Clerc-Schwarzenbach, F.M., Selleri, G., Neri, M., Toth, E., van Meerveld, I., Seibert, J., 2024. HESS Opinions: A few camels or a whole caravan? *EGUsphere* 1–29. <https://doi.org/10.5194/egusphere-2024-864>
- Cleveland, R.B., Cleveland, W.S., McRae, J.E., Terpenning, I., 1990. STL: A seasonal-trend decomposition. *Journal of Official Statistics* 6, 3–73.
- Coulibaly, P., Baldwin, C.K., 2005. Nonstationary hydrological time series forecasting using nonlinear dynamic methods. *Journal of Hydrology* 307, 164–174. <https://doi.org/10.1016/j.jhydrol.2004.10.008>
- Cui, Z., Zhou, Y., Guo, S., Wang, J., Xu, C.-Y., 2022. Effective improvement of multi-step-ahead flood forecasting accuracy through encoder-decoder with an exogenous input structure. *Journal of Hydrology* 609, 127764. <https://doi.org/10.1016/j.jhydrol.2022.127764>

Dangerfield, B.J., Morris, J.S., 1992. Top-down or bottom-up: Aggregate versus disaggregate extrapolations. *International Journal of Forecasting* 8, 233–241. [https://doi.org/10.1016/0169-2070\(92\)90121-O](https://doi.org/10.1016/0169-2070(92)90121-O)

De Livera, A.M., Hyndman, R.J., Snyder, R.D., 2011. Forecasting Time Series With Complex Seasonal Patterns Using Exponential Smoothing. *Journal of the American Statistical Association* 106, 1513–1527.

Delaney, C.J., Hartman, R.K., Mendoza, J., Dettinger, M., Delle Monache, L., Jasperse, J., Ralph, F.M., Talbot, C., Brown, J., Reynolds, D., Evett, S., 2020. Forecast Informed Reservoir Operations Using Ensemble Streamflow Predictions for a Multipurpose Reservoir in Northern California. *Water Resources Research* 56, e2019WR026604. <https://doi.org/10.1029/2019WR026604>

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. <https://doi.org/10.48550/arXiv.1810.04805>

Diebold, F.X., Mariano, R.S., 2002. Comparing Predictive Accuracy. *Journal of Business & Economic Statistics* 20, 134–144. <https://doi.org/10.1198/073500102753410444>

Diks, C.G.H., Vrugt, J.A., 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. *Stoch Environ Res Risk Assess* 24, 809–820. <https://doi.org/10.1007/s00477-010-0378-z>

Dinh, L., Krueger, D., Bengio, Y., 2015. NICE: Non-linear Independent Components Estimation. <https://doi.org/10.48550/arXiv.1410.8516>

Dinh, L., Sohl-Dickstein, J., Bengio, S., 2017. Density estimation using Real NVP. <https://doi.org/10.48550/arXiv.1605.08803>

Dravitzki, S., McGregor, J., 2011. Predictability of heavy precipitation in the Waikato River basin of New Zealand. *Monthly Weather Review* 139, 2184–2197. <https://doi.org/10.1175/2010mwr3137.1>

Dumas, J., Wehenkel, A., Lanaspeze, D., Cornélusse, B., Sutura, A., 2022. A deep generative model for probabilistic energy forecasting in power systems: normalizing flows. *Applied Energy* 305, 117871. <https://doi.org/10.1016/j.apenergy.2021.117871>

Duque, L.-F., O’Connell, E., O’Donnell, G., 2023. A Monte Carlo simulation and sensitivity analysis framework demonstrating the advantages of probabilistic forecasting over deterministic

forecasting in terms of flood warning reliability. *Journal of Hydrology* 619, 129340. <https://doi.org/10.1016/j.jhydrol.2023.129340>

Elman, J.L., 1990. Finding Structure in Time. *Cognitive Science* 14, 179–211. https://doi.org/10.1207/s15516709cog1402_1

Fanfarillo, A., Roozitalab, B., Hu, W., Cervone, G., 2021. Probabilistic forecasting using deep generative models. *Geoinformatica* 25, 127–147. <https://doi.org/10.1007/s10707-020-00425-8>

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, e2019WR026793. <https://doi.org/10.1029/2019WR026793>

Field, C.B., Barros, V.R., Mastrandrea, M.D., Mach, K.J., Abdrabo, M.A.K., Adger, N., 2014. Summary for policymakers, in: *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, pp. 1–32.

Fonzo, T., Girolimetto, D., 2021. Cross-temporal forecast reconciliation: Optimal combination method and heuristic alternatives. *International Journal of Forecasting*. <https://doi.org/10.1016/j.ijforecast.2021.08.004>

Forghani, M., Qian, Y., Lee, J., Farthing, M., Hesser, T., Kitanidis, P.K., Darve, E.F., 2022a. Variational encoder geostatistical analysis (VEGAS) with an application to large scale riverine bathymetry. *Advances in Water Resources* 170, 104323. <https://doi.org/10.1016/j.advwatres.2022.104323>

Forghani, M., Qian, Y., Lee, J., Farthing, M., Hesser, T., Kitanidis, P.K., Darve, E.F., 2022b. Machine Learning and Projection-Based Model Reduction in Hydrology and Geosciences, in: *Knowledge Guided Machine Learning*. Chapman and Hall/CRC.

Forghani, M., Qian, Y., Lee, J., Farthing, M.W., Hesser, T., Kitanidis, P.K., Darve, E.F., 2021. Application of deep learning to large scale riverine flow velocity estimation. *Stoch Environ Res Risk Assess* 35, 1069–1088. <https://doi.org/10.1007/s00477-021-01988-0>

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning.

Gauch, M., Kratzert, F., Klotz, D., Nearing, G., Lin, J., Hochreiter, S., 2021. Rainfall–runoff prediction at multiple timescales with a single Long Short-Term Memory network. *Hydrology and Earth System Sciences* 25, 2045–2062. <https://doi.org/10.5194/hess-25-2045-2021>

Ge, L., Liao, W., Wang, S., Bak-Jensen, B., Pillai, J.R., 2020. Modeling Daily Load Profiles of Distribution Network for Scenario Generation Using Flow-Based Generative Network. *IEEE Access* 8, 77587–77597. <https://doi.org/10.1109/ACCESS.2020.2989350>

Germain, M., Gregor, K., Murray, I., Larochelle, H., 2015. MADE: Masked Autoencoder for Distribution Estimation. <https://doi.org/10.48550/arXiv.1502.03509>

Ghobadi, F., Kang, D., 2022. Multi-Step Ahead Probabilistic Forecasting of Daily Streamflow Using Bayesian Deep Learning: A Multiple Case Study. *Water* 14, 3672. <https://doi.org/10.3390/w14223672>

Gneiting, T., Raftery, A.E., 2007. Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association* 102, 359–378. <https://doi.org/10.1198/016214506000001437>

Goodarzi, L., Banihabib, M.E., Roozbahani, A., 2019. A decision-making model for flood warning system based on ensemble forecasts. *Journal of Hydrology* 573, 207–219. <https://doi.org/10.1016/j.jhydrol.2019.03.040>

Goswami, P., Mohapatra, G.N., 2014. A comparative evaluation of impact of domain size and parameterization scheme on simulation of tropical cyclones in the Bay of Bengal. *Journal of Geophysical Research: Atmospheres* 119, 10–22. <https://doi.org/10.1002/2013jd020592>

Gouda, K.C., Nahak, S., Goswami, P., 2020. Deterministic Seasonal Quantitative Precipitation Forecasts: Benchmark Skill with a GCM. *Pure Appl. Geophys.* 177, 4443–4456. <https://doi.org/10.1007/s00024-020-02463-7>

Green, K.C., Armstrong, J.S., Soon, W., 2009. Validity of climate change forecasting for public policy decision making. *International Journal of Forecasting* 25, 826–832. <https://doi.org/10.1016/j.ijforecast.2009.05.011>

Greenberg, C.H., Goodrick, S., Austin, J.D., Parresol, B.R., 2015. Hydroregime Prediction Models for Ephemeral Groundwater-Driven Sinkhole Wetlands: a Planning Tool for Climate Change and Amphibian Conservation. *Wetlands* 35, 899–911. <https://doi.org/10.1007/s13157-015-0680-0>

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, 80–91. <https://doi.org/10.1016/j.jhydrol.2009.08.003>

Han, H., Choi, C., Jung, J., Kim, H.S., 2021. Deep Learning with Long Short Term Memory Based Sequence-to-Sequence Model for Rainfall-Runoff Simulation. *Water* 13, 437. <https://doi.org/10.3390/w13040437>

Han, H., Morrison, R.R., 2022. Improved runoff forecasting performance through error predictions using a deep-learning approach. *J. Hydrol* 608, 127653.

Hao, Y., Baik, J., Tran, H., Choi, M., 2022. Quantification of the effect of hydrological drivers on actual evapotranspiration using the Bayesian model averaging approach for various landscapes over Northeast Asia. *Journal of Hydrology* 607, 127543. <https://doi.org/10.1016/j.jhydrol.2022.127543>

Harris, L., McRae, A.T.T., Chantry, M., Dueben, P.D., Palmer, T.N., 2022. A Generative Deep Learning Approach to Stochastic Downscaling of Precipitation Forecasts. *Journal of Advances in Modeling Earth Systems* 14, e2022MS003120. <https://doi.org/10.1029/2022MS003120>

Harvey, D., Leybourne, S., Newbold, P., 1997. Testing the equality of prediction mean squared errors. *International Journal of Forecasting* 13, 281–291. [https://doi.org/10.1016/s0169-2070\(96\)00719-4](https://doi.org/10.1016/s0169-2070(96)00719-4)

Hauswirth, S.M., Bierkens, M.F.P., Beijk, V., Wanders, N., 2023. The suitability of a seasonal ensemble hybrid framework including data-driven approaches for hydrological forecasting. *Hydrology and Earth System Sciences* 27, 501–517. <https://doi.org/10.5194/hess-27-501-2023>

Hewamalage, H., Bergmeir, C., Bandara, K., 2021. Recurrent neural networks for time series forecasting. Current status and future directions. *International Journal of Forecasting* 37, 388–427. <https://doi.org/10.1016/j.ijforecast.2020.06.008>

Higgins, R.W., Mo, K.C., Schubert, S.D., 1996. The moisture budget of the central United States in spring as evaluated in the NCEP/ NCAR and the NASA/DAO reanalyses. *Monthly Weather Review* 124, 939–963.

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. *Neural Computation* 9, 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>

Höge, M., Guthke, A., Nowak, W., 2019. The hydrologist's guide to Bayesian model selection, averaging and combination. *Journal of Hydrology* 572, 96–107. <https://doi.org/10.1016/j.jhydrol.2019.01.072>

Hong, T., Xie, J., Black, J., 2019. Global energy forecasting competition 2017: Hierarchical probabilistic load forecasting. *International Journal of Forecasting* 35, 1389–1399. <https://doi.org/10.1016/j.ijforecast.2019.02.006>

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. *Neural Networks* 2, 359–366. [https://doi.org/10.1016/0893-6080\(89\)90020-8](https://doi.org/10.1016/0893-6080(89)90020-8)

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, 1543. <https://doi.org/10.3390/w10111543>

Huang, X., Li, Q., Tai, Y., Chen, Z., Liu, J., Shi, J., Liu, W., 2022. Time series forecasting for hourly photovoltaic power using conditional generative adversarial network and Bi-LSTM. *Energy* 246, 123403. <https://doi.org/10.1016/j.energy.2022.123403>

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, 725–741. <https://doi.org/10.1175/2008JAMC1979.1>

Hyndman, R., Koehler, A.B., Ord, J.K., Snyder, R.D., 2008. Forecasting with exponential smoothing: The state space approach. Springer Science & Business Media.

Hyndman, Rob J., Ahmed, R.A., Athanasopoulos, G., Shang, H.L., 2011. Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis* 55, 2579–2589. <https://doi.org/10.1016/j.csda.2011.03.006>

Hyndman, R.J., Athanasopoulos, G., 2018. Forecasting: Principles and practice. OTexts.

Hyndman, R.J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., 2022. forecast: Forecasting functions for time series and linear models.

Hyndman, R.J., Khandakar, Y., 2008. Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software* 27, 1–22. <https://doi.org/10.18637/jss.v027.i03>

Hyndman, R.J., Lee, A.J., Wang, E., 2016. Fast computation of reconciled forecasts for hierarchical and grouped time series. *Computational Statistics & Data Analysis* 97, 16–32. <https://doi.org/10.1016/j.csda.2015.11.007>

Jahangir, M.S., Quilty, J., 2024. Generative deep learning for probabilistic streamflow forecasting: Conditional variational auto-encoder. *Journal of Hydrology* 629, 130498. <https://doi.org/10.1016/j.jhydrol.2023.130498>

Jahangir, M.S., Quilty, J., 2022. Temporal Hierarchical Reconciliation for Consistent Water Resources Forecasting Across Multiple Timescales: An Application to Precipitation Forecasting. *Water Resources Research* 58, e2021WR031862. <https://doi.org/10.1029/2021WR031862>

Jahangir, M.S., You, J., Quilty, J., 2023. A quantile-based encoder-decoder framework for multi-step ahead runoff forecasting. *Journal of Hydrology* 619, 129269. <https://doi.org/10.1016/j.jhydrol.2023.129269>

Jee, J.B., Kim, S., 2017. Sensitivity study on high-resolution WRF precipitation forecast for a heavy rainfall event. *Atmosphere* 8, 96. <https://doi.org/10.3390/atmos8060096>

Jeon, J., Panagiotelis, A., Petropoulos, F., 2019. Probabilistic forecast reconciliation with applications to wind power and electric load. *European Journal of Operational Research* 279, 364–379. <https://doi.org/10.1016/j.ejor.2019.05.020>

Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K., 1999. An Introduction to Variational Methods for Graphical Models. *Machine Learning* 37, 183–233. <https://doi.org/10.1023/A:1007665907178>

Kang, X., Kokkinaki, A., Kitanidis, P.K., Shi, X., Lee, J., Mo, S., Wu, J., 2021. Hydrogeophysical Characterization of Nonstationary DNAPL Source Zones by Integrating a Convolutional Variational Autoencoder and Ensemble Smoother. *Water Resources Research* 57, e2020WR028538. <https://doi.org/10.1029/2020WR028538>

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>

Kelly, K.S., Krzysztofowicz, R., 2000. Precipitation uncertainty processor for probabilistic river stage forecasting. *Water Resources Research* 36, 2643–2653. <https://doi.org/10.1029/2000WR900061>

Ketabchi, H., Jahangir, M.S., 2021. Influence of aquifer heterogeneity on sea level rise-induced seawater intrusion: A probabilistic approach. *Journal of Contaminant Hydrology* 236, 103753. <https://doi.org/10.1016/j.jconhyd.2020.103753>

Ketabchi, H., Jahangir, M.S., 2019. Probabilistic numerical assessment of seawater intrusion overshoot in heterogeneous coastal aquifers. *Stoch Environ Res Risk Assess* 33, 1951–1968. <https://doi.org/10.1007/s00477-019-01744-5>

Kingma, D.P., Salimans, T., Jozefowicz, R., Chen, X., Sutskever, I., Welling, M., 2017. Improving Variational Inference with Inverse Autoregressive Flow. <https://doi.org/10.48550/arXiv.1606.04934>

Kingma, D.P., Welling, M., 2022. Auto-Encoding Variational Bayes. <https://doi.org/10.48550/arXiv.1312.6114>

Kling, H., Fuchs, M., Paulin, M., 2012. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *Journal of Hydrology* 424–425, 264–277. <https://doi.org/10.1016/j.jhydrol.2012.01.011>

Klotz, D., Kratzert, F., Gauch, M., Keefe Sampson, A., Brandstetter, J., Klambauer, G., Hochreiter, S., Nearing, G., 2022. Uncertainty estimation with deep learning for rainfall–runoff modeling. *Hydrology and Earth System Sciences* 26, 1673–1693. <https://doi.org/10.5194/hess-26-1673-2022>

Knoben, W.J.M., Freer, J.E., Fowler, K.J.A., Peel, M.C., Woods, R.A., 2019. Modular Assessment of Rainfall–Runoff Models Toolbox (MARRMoT) v1.2: an open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations. *Geoscientific Model Development* 12, 2463–2480. <https://doi.org/10.5194/gmd-12-2463-2019>

Kobold, M., Sušelj, K., 2005. Precipitation forecasts and their uncertainty as input into hydrological models. *Hydrology and Earth System Sciences* 9, 322–332. <https://doi.org/10.5194/hess-9-322-2005>

Kourentzes, N., Athanasopoulos, G., 2019. Cross-temporal coherent forecasts for Australian tourism. *Annals of Tourism Research* 75, 393–409. <https://doi.org/10.1016/j.annals.2019.02.001>

Kourentzes, Nikolaos, Athanasopoulos, G., 2019. Cross-temporal coherent forecasts for Australian tourism. *Annals of Tourism Research* 75, 393–409. <https://doi.org/10.1016/j.annals.2019.02.001>

Kourentzes, N., Petropoulos, F., 2016. Forecasting with multivariate temporal aggregation: The case of promotional modelling. *International Journal of Production Economics*, SI: ISIR 2014 181, 145–153. <https://doi.org/10.1016/j.ijpe.2015.09.011>

Kourentzes, N., Rostami-Tabar, B., Barrow, D.K., 2017. Demand forecasting by temporal aggregation: Using optimal or multiple aggregation levels? *Journal of Business Research* 78, 1–9. <https://doi.org/10.1016/j.jbusres.2017.04.016>

Koutsoyiannis, D., Onof, C., Wheeler, H.S., 2003. Multivariate rainfall disaggregation at a fine timescale. *Water Resources Research* 39. <https://doi.org/10.1029/2002wr001600>

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, 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, 5089–5110. <https://doi.org/10.5194/hess-23-5089-2019>

Kratzert, F., Nearing, G., Addor, N., Erickson, T., Gauch, M., Gilon, O., Gudmundsson, L., Hassidim, A., Klotz, D., Nevo, S., Shalev, G., Matias, Y., 2023. Caravan - A global community dataset for large-sample hydrology. *Sci Data* 10, 61. <https://doi.org/10.1038/s41597-023-01975-w>

Kullback, S., Leibler, R.A., 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* 22, 79–86.

Kumar, D.N., Lall, U., Petersen, M.R., 2000. Multisite disaggregation of monthly to daily streamflow. *Water Resources Research* 36, 1823–1833. <https://doi.org/10.1029/2000wr900049>

Kwiatkowski, D., Phillips, P.C., Schmidt, P., Shin, Y., 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics* 54, 159–178.

Le, X.-H., Ho, H.V., Lee, G., Jung, S., 2019. Application of Long Short-Term Memory (LSTM) Neural Network for Flood Forecasting. *Water* 11, 1387. <https://doi.org/10.3390/w11071387>

Li, D., Marshall, L., Liang, Z., Sharma, A., Zhou, Y., 2021. Bayesian LSTM With Stochastic Variational Inference for Estimating Model Uncertainty in Process-Based Hydrological Models. *Water Resources Research* 57, e2021WR029772. <https://doi.org/10.1029/2021WR029772>

Li, J., Yuan, D., Liu, J., Jiang, Y., Chen, Y., Hsu, K.L., Sorooshian, S., 2019. Predicting floods in a large karst river basin by coupling PERSIANN-CCS QPEs with a physically based distributed hydrological model. *Hydrology and Earth System Sciences* 23, 1505–1532. <https://doi.org/10.5194/hess-23-1505-2019>

Li, Y., Lu, G., Wu, Z., He, H., Shi, J., Ma, Y., Weng, S., 2016. Evaluation of optimized WRF precipitation forecast over a complex topography region during flood season. *Atmosphere* 7, 145. <https://doi.org/10.3390/atmos7110145>

Liu, D., Mishra, A.K., Yu, Z., Yang, C., Konapala, G., Vu, T., 2017. Performance of SMAP, AMSR-E and LAI for weekly agricultural drought forecasting over continental United States. *Journal of Hydrology* 553, 88–104. <https://doi.org/10.1016/j.jhydrol.2017.07.049>

Liu, Z., Zhou, P., Chen, X., Guan, Y., 2015. A multivariate conditional model for streamflow prediction and spatial precipitation refinement. *Journal of Geophysical Research: Atmospheres* 120, 10,116–10,129. <https://doi.org/10.1002/2015JD023787>

Lombardo, F., Volpi, E., Koutsoyiannis, D., Serinaldi, F., 2017. A theoretically consistent stochastic cascade for temporal disaggregation of intermittent rainfall. *Water Resources Research* 53, 4586–4605. <https://doi.org/10.1002/2017wr020529>

Lu, D., Konapala, G., Painter, S.L., Kao, S.-C., Gangrade, S., 2021. Streamflow Simulation in Data-Scarce Basins Using Bayesian and Physics-Informed Machine Learning Models. *Journal of Hydrometeorology* 22, 1421–1438. <https://doi.org/10.1175/JHM-D-20-0082.1>

Lu, K., Wang, L., 2011. A novel nonlinear combination model based on support vector machine for rainfall prediction, in: 2011 Fourth International Joint Conference on Computational Sciences and Optimization. IEEE, pp. 1343–1346.

Luo, L., Wood, E.F., 2008. Use of Bayesian Merging Techniques in a Multimodel Seasonal Hydrologic Ensemble Prediction System for the Eastern United States. *Journal of Hydrometeorology* 9, 866–884. <https://doi.org/10.1175/2008JHM980.1>

Luo, X., Nadiga, B.T., Ren, Y., Park, J.H., Xu, W., Yoo, S., 2022. A Bayesian Deep Learning Approach to Near-Term Climate Prediction. *J Adv Model Earth Syst* 14. <https://doi.org/10.1029/2022MS003058>

Lv, N., Liang, X., Chen, C., Zhou, Y., Li, J., Wei, H., Wang, H., 2020. A long Short-Term memory cyclic model with mutual information for hydrology forecasting: A Case study in the xixian basin. *Advances in Water Resources* 141, 103622. <https://doi.org/10.1016/j.advwatres.2020.103622>

Mancuso, P., Piccialli, V., Sudoso, A.M., 2021. A machine learning approach for forecasting hierarchical time series. *Expert Systems with Applications* 182, 115102. <https://doi.org/10.1016/j.eswa.2021.115102>

McCuen, R.H., Knight, Z., Cutter, A.G., 2006. Evaluation of the Nash–Sutcliffe Efficiency Index. *Journal of Hydrologic Engineering* 11, 597–602. [https://doi.org/10.1061/\(ASCE\)1084-0699\(2006\)11:6\(597\)](https://doi.org/10.1061/(ASCE)1084-0699(2006)11:6(597))

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, 5669–5683. <https://doi.org/10.5194/hess-26-5669-2022>

Mehr, A.D., Kahya, E., Olyaie, E., 2013. Streamflow prediction using linear genetic programming in comparison with a neuro-wavelet technique. *J. Hydrol* 505, 240–249.

Moges, E., Demissie, Y., Larsen, L., Yassin, F., 2020. Sources of hydrological model uncertainties and advances in their analysis. *Water* 13, 28.

Montanari, A., Koutsoyiannis, D., 2012. A blueprint for process-based modeling of uncertain hydrological systems. *Water Resources Research* 48. <https://doi.org/10.1029/2011WR011412>

Moriasi, D.N., Arnold, J.G., Liew, M.W., Bingner, R.L., Harmel, R.D., Veith, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE* 50, 885–900. <https://doi.org/10.13031/2013.23153>

Myronidis, D., Ioannou, K., Fotakis, D., Dörflinger, G., 2018. Streamflow and Hydrological Drought Trend Analysis and Forecasting in Cyprus. *Water Resour Manage* 32, 1759–1776. <https://doi.org/10.1007/s11269-018-1902-z>

Nai, C., Liu, X., Tang, Q., Liu, L., Sun, S., Gaffney, P.P.J., 2024. A Novel Strategy for Automatic Selection of Cross-Basin Data to Improve Local Machine Learning-Based Runoff Models. *Water Resources Research* 60, e2023WR035051. <https://doi.org/10.1029/2023WR035051>

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology* 10, 282–290. [https://doi.org/10.1016/0022-1694\(70\)90255-6](https://doi.org/10.1016/0022-1694(70)90255-6)

Navarro-Ortega, A., Acuña, V., Bellin, A., Burek, P., Cassiani, G., Choukr-Allah, R., Dolédec, S., Elozegi, A., Ferrari, F., Ginebreda, A., Grathwohl, P., Jones, C., Rault, P.K., Kok, K., Koundouri, P., Ludwig, R.P., Merz, R., Milacic, R., Muñoz, I., Nikulin, G., Paniconi, C., Paunović, M., Petrovic, M., Sabater, L., Sabater, S., Skoulikidis, N.Th., Slob, A., Teutsch, G., Voulvoulis, N., Barceló, D., 2015. Managing the effects of multiple stressors on aquatic ecosystems under water scarcity. The GLOBAQUA project. *Science of The Total Environment, Towards a better understanding of the links between stressors, hazard assessment and ecosystem services under water scarcity* 503–504, 3–9. <https://doi.org/10.1016/j.scitotenv.2014.06.081>

Nearing, G., Cohen, D., Dube, V., Gauch, M., Gilon, O., Harrigan, S., Hassidim, A., Klotz, D., Kratzert, F., Metzger, A., Nevo, S., Pappenberger, F., Prudhomme, C., Shalev, G., Shenzis, S., Tekalign, T.Y., Weitzner, D., Matias, Y., 2024. Global prediction of extreme floods in ungauged watersheds. *Nature* 627, 559–563. <https://doi.org/10.1038/s41586-024-07145-1>

Nearing, G.S., Ruddell, B.L., Clark, M.P., Nijssen, B., Peters-Lidard, C., 2018. Benchmarking and Process Diagnostics of Land Models. *Journal of Hydrometeorology* 19, 1835–1852. <https://doi.org/10.1175/JHM-D-17-0209.1>

Nevo, S., Morin, E., Gerzi Rosenthal, A., Metzger, A., Barshai, C., Weitzner, D., Voloshin, D., Kratzert, F., Elidan, G., Dror, G., Begelman, G., Nearing, G., Shalev, G., Noga, H., Shavitt, I., Yuklea, L., Royz, M., Giladi, N., Peled Levi, N., Reich, O., Gilon, O., Maor, R., Timnat, S., Shechter, T., Anisimov, V., Gigi, Y., Levin, Y., Moshe, Z., Ben-Haim, Z., Hassidim, A., Matias, Y., 2022. Flood forecasting with machine learning models in an operational framework. *Hydrology and Earth System Sciences* 26, 4013–4032. <https://doi.org/10.5194/hess-26-4013-2022>

Newman, A.J., Clark, M.P., Sampson, K., Wood, A., Hay, L.E., Bock, A., Viger, R.J., Blodgett, D., Brekke, L., Arnold, J.R., Hopson, T., Duan, Q., 2015. Development of a large-sample watershed-scale

hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences* 19, 209–223. <https://doi.org/10.5194/hess-19-209-2015>

Ni, L., Wang, D., Singh, V.P., Wu, J., Wang, Y., Tao, Y., Zhang, J., 2020. Streamflow and rainfall forecasting by two long short-term memory-based models. *Journal of Hydrology* 583, 124296. <https://doi.org/10.1016/j.jhydrol.2019.124296>

Nystrup, P., Lindström, E., Pinson, P., Madsen, H., 2020. Temporal hierarchies with autocorrelation for load forecasting. *European Journal of Operational Research* 280, 876–888. <https://doi.org/10.1016/j.ejor.2019.07.061>

Olivares, K.G., Meetei, O.N., Ma, R., Reddy, R., Cao, M., Dicker, L., 2024. Probabilistic hierarchical forecasting with deep Poisson mixtures. *International Journal of Forecasting* 40, 470–489. <https://doi.org/10.1016/j.ijforecast.2023.04.007>

Oord, A., Li, Y., Babuschkin, I., Simonyan, K., Vinyals, O., Kavukcuoglu, K., Driessche, G., Lockhart, E., Cobo, L., Stimberg, F., Casagrande, N., Grewe, D., Noury, S., Dieleman, S., Elsen, E., Kalchbrenner, N., Zen, H., Graves, A., King, H., Walters, T., Belov, D., Hassabis, D., 2018. Parallel WaveNet: Fast High-Fidelity Speech Synthesis, in: *Proceedings of the 35th International Conference on Machine Learning*. Presented at the International Conference on Machine Learning, PMLR, pp. 3918–3926.

Pan, B., Hsu, K., AghaKouchak, A., Sorooshian, S., 2019. Improving precipitation estimation using convolutional neural network. *Water Resources Research* 55, 2301–2321. <https://doi.org/10.1029/2018wr024090>

Papacharalampous, G., Tyralis, H., Koutsoyiannis, D., 2019. Comparison of stochastic and machine learning methods for multi-step ahead forecasting of hydrological processes. *Stochastic Environmental Research and Risk Assessment* 33, 481–514.

Papacharalampous, G., Tyralis, H., Koutsoyiannis, D., 2018a. Univariate time series forecasting of temperature and precipitation with a focus on machine learning algorithms: A multiple-case study from Greece. *Water Resources Management* 32, 5207–5239. <https://doi.org/10.1007/s11269-018-2155-6>

Papacharalampous, G., Tyralis, H., Koutsoyiannis, D., 2018b. Predictability of monthly temperature and precipitation using automatic time series forecasting methods. *Acta Geophysica* 66, 807–831. <https://doi.org/10.1007/s11600-018-0120-7>

Papacharalampous, G., Tyralis, H., Koutsoyiannis, D., Montanari, A., 2020. Quantification of predictive uncertainty in hydrological modelling by harnessing the wisdom of the crowd: A large-sample experiment at monthly timescale. *Advances in Water Resources* 136, 103470. <https://doi.org/10.1016/j.advwatres.2019.103470>

Papacharalampous, G., Tyralis, H., Papalexiou, S.M., Langousis, A., Khatami, S., Volpi, E., Grimaldi, S., 2021. Global-scale massive feature extraction from monthly hydroclimatic time series: Statistical characterizations, spatial patterns and hydrological similarity. *Science of The Total Environment* 767, 144612. <https://doi.org/10.1016/j.scitotenv.2020.144612>

Papamakarios, G., Pavlakou, T., Murray, I., 2018. Masked Autoregressive Flow for Density Estimation. <https://doi.org/10.48550/arXiv.1705.07057>

Pappenberger, F., Ramos, M.H., Cloke, H.L., Wetterhall, F., Alfieri, L., Bogner, K., Mueller, A., Salamon, P., 2015. How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *Journal of Hydrology* 522, 697–713. <https://doi.org/10.1016/j.jhydrol.2015.01.024>

Partington, D., Brunner, P., Simmons, C.T., Werner, A.D., Therrien, R., Maier, H.R., Dandy, G.C., 2012. Evaluation of outputs from automated baseflow separation methods against simulated baseflow from a physically based, surface water- groundwater flow model. *J. Hydrol* 458, 28–39.

Peters-Lidard, C.D., Hossain, F., Leung, L.R., McDowell, N., Rodell, M., Tapiador, F.J., 2019. 100 years of progress in hydrology. *Meteorological Monographs* 59, 25–31.

Petropoulos, F., Kourentzes, N., 2015. Forecast combinations for intermittent demand. *Journal of the Operational Research Society* 66, 914–924. <https://doi.org/10.1057/jors.2014.62>

Punia, S., Singh, S.P., Madaan, J.K., 2020. A cross-temporal hierarchical framework and deep learning for supply chain forecasting. *Computers & Industrial Engineering* 149, 106796. <https://doi.org/10.1016/j.cie.2020.106796>

Qi, W., Liu, J., Yang, H., Sweetapple, C., 2018. An ensemble-based dynamic Bayesian averaging approach for discharge simulations using multiple global precipitation products and hydrological models. *Journal of Hydrology* 558, 405–420. <https://doi.org/10.1016/j.jhydrol.2018.01.026>

Quilty, J., Adamowski, J., 2020. A stochastic wavelet-based data-driven framework for forecasting uncertain multiscale hydrological and water resources processes. *Environ. Model. Softw* 130, 104718.

Quilty, J., Jahangir, M.S., You, J., Hughes, H., Hah, D., Tzoganakis, I., 2023. Bayesian extreme learning machines for hydrological prediction uncertainty. *Journal of Hydrology* 626, 130138. <https://doi.org/10.1016/j.jhydrol.2023.130138>

Rangapuram, S.S., Werner, L.D., Benidis, K., Mercado, P., Gasthaus, J., Januschowski, T., 2021. End-to-End Learning of Coherent Probabilistic Forecasts for Hierarchical Time Series, in: *Proceedings of the 38th International Conference on Machine Learning*. Presented at the International Conference on Machine Learning, PMLR, pp. 8832–8843.

Rasiya Koya, S., Roy, T., 2024. Temporal Fusion Transformers for streamflow Prediction: Value of combining attention with recurrence. *Journal of Hydrology* 637, 131301. <https://doi.org/10.1016/j.jhydrol.2024.131301>

Rasul, K., Seward, C., Schuster, I., Vollgraf, R., 2021. Autoregressive Denoising Diffusion Models for Multivariate Probabilistic Time Series Forecasting, in: *Proceedings of the 38th International Conference on Machine Learning*. Presented at the International Conference on Machine Learning, PMLR, pp. 8857–8868.

Rostami-Tabar, B., Babai, M.Z., Syntetos, A., Ducq, Y., 2014. A note on the forecast performance of temporal aggregation. *Naval Research Logistics* 61, 489–500. <https://doi.org/10.1002/nav.21598>

Rostami-Tabar, B., Babai, M.Z., Syntetos, A., Ducq, Y., 2013. Demand forecasting by temporal aggregation. *Naval Research Logistics* 60, 479–498. <https://doi.org/10.1002/nav.21546>

Ryu, Y., Baldocchi, D.D., Black, T.A., Detto, M., Law, B.E., Leuning, R., 2012. On the temporal upscaling of evapotranspiration from instantaneous remote sensing measurements to 8-day mean daily-sums. *Agricultural and Forest Meteorology* 152, 212–222. <https://doi.org/10.1016/j.agrformet.2011.09.010>

Sabzipour, B., Arsenault, R., Troin, M., Martel, J.-L., Brissette, F., Brunet, F., Mai, J., 2023. Comparing a long short-term memory (LSTM) neural network with a physically-based hydrological

model for streamflow forecasting over a Canadian catchment. *Journal of Hydrology* 627, 130380. <https://doi.org/10.1016/j.jhydrol.2023.130380>

Salazar, A.A., Zheng, J., Che, Y., Xiao, F., 2022. Deep generative model for probabilistic wind speed and wind power estimation at a wind farm. *Energy Science & Engineering* 10, 1855–1873. <https://doi.org/10.1002/ese3.1086>

Seeger, M., 2004. Gaussian processes for machine learning. *Int. J. Neural Syst* 14, 69–106.

Shang, Y., Xu, Y., Shang, L., Fan, Q., Wang, Y., Liu, Z., 2019. A method of direct, real-time forecasting of downstream water levels via hydropower station reregulation: A case study from Gezhouba Hydropower Plant, China. *Journal of Hydrology* 573, 895–907. <https://doi.org/10.1016/j.jhydrol.2019.04.017>

Sharma, T.C., Panu, U.S., 2014. Modeling of hydrological drought durations and magnitudes: Experiences on Canadian streamflows. *Journal of Hydrology: Regional Studies* 1, 92–106. <https://doi.org/10.1016/j.ejrh.2014.06.006>

Sheather, S.J., 2004. Density estimation. *Statistical Science* 19, 588–597. <https://doi.org/10.1214/088342304000000297>

Silvestrini, A., Veredas, D., 2008. Temporal aggregation of univariate and multivariate time series models: A survey. *Journal of Economic Surveys* 22, 458–497. <https://doi.org/10.1111/j.1467-6419.2007.00538.x>

Solomatine, D.P., Ostfeld, A., 2008. Data-driven modelling: some past experiences and new approaches. *Journal of Hydroinformatics* 10, 3–22. <https://doi.org/10.2166/hydro.2008.015>

Sønderby, C.K., Espeholt, L., Heek, J., Dehghani, M., Oliver, A., Salimans, T., 2020. MetNet: A neural weather model for precipitation forecasting.

Spiliotis, E., Abolghasemi, M., Hyndman, R.J., Petropoulos, F., Assimakopoulos, V., 2021. Hierarchical forecast reconciliation with machine learning. *Applied Soft Computing* 112, 107756. <https://doi.org/10.1016/j.asoc.2021.107756>

Spiliotis, E., Petropoulos, F., Kourentzes, N., Assimakopoulos, V., 2020. Cross-temporal aggregation: Improving the forecast accuracy of hierarchical electricity consumption. *Applied Energy* 261, 114339. <https://doi.org/10.1016/j.apenergy.2019.114339>

Sun, A.Y., Jiang, P., Yang, Z.-L., Xie, Y., Chen, X., 2022. A graph neural network (GNN) approach to basin-scale river network learning: the role of physics-based connectivity and data fusion. *Hydrology and Earth System Sciences* 26, 5163–5184. <https://doi.org/10.5194/hess-26-5163-2022>

Taieb, S.B., Taylor, J.W., Hyndman, R.J., 2017. Coherent Probabilistic Forecasts for Hierarchical Time Series, in: *Proceedings of the 34th International Conference on Machine Learning*. Presented at the International Conference on Machine Learning, PMLR, pp. 3348–3357.

Taylor, S.J., Letham, B., 2018. Forecasting at scale. *The American Statistician* 72, 37–45. <https://doi.org/10.1080/00031305.2017.138>

Thiboult, A., Anctil, F., Boucher, M.-A., 2016. Accounting for three sources of uncertainty in ensemble hydrological forecasting. *Hydrology and Earth System Sciences* 20, 1809–1825. <https://doi.org/10.5194/hess-20-1809-2016>

Tian, D., Wood, E.F., Yuan, X., 2017. CFSv2-based sub-seasonal precipitation and temperature forecast skill over the contiguous United States. *Hydrology and Earth System Sciences* 21, 1477–1490. <https://doi.org/10.5194/hess-21-1477-2017>

Tiao, G.C., 1972. Asymptotic behaviour of temporal aggregates of time series. *Biometrika* 59, 525–531.

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, e2020WR028392. <https://doi.org/10.1029/2020WR028392>

Tsoukalas, I., Efstratiadis, A., Makropoulos, C., 2019. Building a puzzle to solve a riddle: A multi-scale disaggregation approach for multivariate stochastic processes with any marginal distribution and correlation structure. *Journal of Hydrology* 575, 354–380. <https://doi.org/10.1016/j.jhydrol.2019.05.017>

Valipour, M., Banihabib, M.E., Behbahani, S.M.R., 2013. Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. *Journal of Hydrology* 476, 433–441. <https://doi.org/10.1016/j.jhydrol.2013.05.017>

van den Hurk, B.J.J.M., Bouwer, L.M., Buontempo, C., Döscher, R., Ercin, E., Hananel, C., Hunink, J.E., Kjellström, E., Klein, B., Manez, M., Pappenberger, F., Pouget, L., Ramos, M.-H., Ward, P.J., Weerts, A.H., Wijngaard, J.B., 2016. Improving predictions and management of hydrological extremes

through climate services: www.imprex.eu. Climate Services 1, 6–11.
<https://doi.org/10.1016/j.cliser.2016.01.001>

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. ukasz, Polosukhin, I., 2017. Attention is All you Need, in: Advances in Neural Information Processing Systems. Curran Associates, Inc.

Velázquez, J.A., Petit, T., Lavoie, A., Boucher, M.-A., Turcotte, R., Fortin, V., Anctil, F., 2009. An evaluation of the Canadian global meteorological ensemble prediction system for short-term hydrological forecasting. Hydrology and Earth System Sciences 13, 2221–2231.
<https://doi.org/10.5194/hess-13-2221-2009>

Villegas, M.A., Pedregal, D.J., 2018. Supply chain decision support systems based on a novel hierarchical forecasting approach. Decision Support Systems 114, 29–36.
<https://doi.org/10.1016/j.dss.2018.08.003>

Vitart, F., 2004. Monthly forecasting at ECMWF. Monthly Weather Review 132, 2761–2779.
<https://doi.org/10.1175/mwr2826.1>

Wang, E., Zhang, Y., Luo, J., Chiew, F.H.S., Wang, Q.J., 2011. Monthly and seasonal streamflow forecasts using rainfall-runoff modeling and historical weather data. Water Resources Research 47.
<https://doi.org/10.1029/2010WR009922>

Wang, H.R., Wang, C., Lin, X., Kang, J., 2014. An improved ARIMA model for precipitation simulations. Nonlinear Processes in Geophysics 21, 1159–1168. <https://doi.org/10.5194/npg-21-1159-2014>

Wang, X., Smith, K., Hyndman, R., 2006. Characteristic-based clustering for time series data. Data Mining and Knowledge Discovery 13, 335–364. <https://doi.org/10.1007/s10618-005-0039-x>

Wang, Y., Gan, D., Sun, M., Zhang, N., Lu, Z., Kang, C., 2019. Probabilistic individual load forecasting using pinball loss guided LSTM. Applied Energy 235, 10–20.
<https://doi.org/10.1016/j.apenergy.2018.10.078>

Werner, K., Jansson, M., Stoica, P., 2008. On Estimation of Covariance Matrices With Kronecker Product Structure. IEEE Transactions on Signal Processing 56, 478–491.
<https://doi.org/10.1109/TSP.2007.907834>

Wickramasuriya, S.L., Athanasopoulos, G., Hyndman, R.J., 2019. Optimal Forecast Reconciliation for Hierarchical and Grouped Time Series Through Trace Minimization. *Journal of the American Statistical Association* 114, 804–819. <https://doi.org/10.1080/01621459.2018.1448825>

Wilbrand, K., Taormina, R., ten Veldhuis, M.-C., Visser, M., Hrachowitz, M., Nuttall, J., Dahm, R., 2023. Predicting streamflow with LSTM networks using global datasets. *Frontiers in Water* 5.

Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., Deng, S.-H., 2019a. Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. *Journal of Electronic Science and Technology* 17, 26–40. <https://doi.org/10.11989/JEST.1674-862X.80904120>

Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., Deng, S.-H., 2019b. Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization. *Journal of Electronic Science and Technology* 17, 26–40. <https://doi.org/10.11989/JEST.1674-862X.80904120>

Xiang, Z., Yan, J., Demir, I., 2020. A Rainfall-Runoff Model With LSTM-Based Sequence-to-Sequence Learning. *Water Resources Research* 56, e2019WR025326. <https://doi.org/10.1029/2019WR025326>

Xie, K., Liu, P., Zhang, J., Han, D., Wang, G., Shen, C., 2021. Physics-guided deep learning for rainfall-runoff modeling by considering extreme events and monotonic relationships. *Journal of Hydrology* 603, 127043. <https://doi.org/10.1016/j.jhydrol.2021.127043>

Xu, Y., Lin, K., Hu, C., Wang, S., Wu, Q., Zhang, L., Ran, G., 2023. Deep transfer learning based on transformer for flood forecasting in data-sparse basins. *Journal of Hydrology* 625, 129956. <https://doi.org/10.1016/j.jhydrol.2023.129956>

Xuan, Y., Cluckie, I.D., Wang, Y., 2009. Uncertainty analysis of hydrological ensemble forecasts in a distributed model utilising short-range rainfall prediction. *Hydrology and Earth System Sciences* 13, 293–303. <https://doi.org/10.5194/hess-13-293-2009>

Yagli, G.M., Yang, D., Srinivasan, D., 2019. Reconciling solar forecasts: Sequential reconciliation. *Solar Energy* 179, 391–397. <https://doi.org/10.1016/j.solener.2018.12.075>

Yan, W., 2012. Toward automatic time-series forecasting using neural networks. *IEEE Transactions on Neural Networks and Learning Systems* 23, 1028–1039. <https://doi.org/10.1109/tnnls.2012.2198074>

- Yang, D., Quan, H., Disfani, V.R., Rodríguez-Gallegos, C.D., 2017. Reconciling solar forecasts: Temporal hierarchy. *Solar Energy* 158, 332–346. <https://doi.org/10.1016/j.solener.2017.09.055>
- Yao, C., Ye, J., He, Z., Bastola, S., Zhang, K., Li, Z., 2019. Evaluation of flood prediction capability of the distributed Grid-Xinjiang model driven by weather research and forecasting precipitation. *Journal of Flood Risk Management* 12, e12544. <https://doi.org/10.1111/jfr3.12544>
- Yeditha, P.K., Kasi, V., Rathinasamy, M., Agarwal, A., 2020. Forecasting of extreme flood events using different satellite precipitation products and wavelet-based machine learning methods. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 30, 063115. <https://doi.org/10.1063/5.0008195>
- Yin, H., Guo, Z., Zhang, X., Chen, J., Zhang, Y., 2022. RR-Former: Rainfall-runoff modeling based on Transformer. *Journal of Hydrology* 609, 127781. <https://doi.org/10.1016/j.jhydrol.2022.127781>
- Yin, H., Zhang, X., Wang, F., Zhang, Y., Xia, R., Jin, J., 2021. Rainfall-runoff modeling using LSTM-based multi-state-vector sequence-to-sequence model. *Journal of Hydrology* 598, 126378. <https://doi.org/10.1016/j.jhydrol.2021.126378>
- Zhang, G., Patuwo, B.E., Hu, M.Y., 1998. Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting* 14, 35–62. [https://doi.org/10.1016/s0169-2070\(97\)00044-7](https://doi.org/10.1016/s0169-2070(97)00044-7)
- Zhang, G.P., 2001. An investigation of neural networks for linear time-series forecasting. *Computers & Operations Research* 28, 1183–1202. [https://doi.org/10.1016/s0305-0548\(00\)00033-2](https://doi.org/10.1016/s0305-0548(00)00033-2)
- Zhang, J., Chen, X., Khan, A., Zhang, Y., Kuang, X., Liang, X., Taccari, M.L., Nuttall, J., 2021. Daily runoff forecasting by deep recursive neural network. *Journal of Hydrology* 596, 126067. <https://doi.org/10.1016/j.jhydrol.2021.126067>
- 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. *J. Hydrol* 530, 137–152.
- Zhou, F., Wang, Z., Chen, D., Zhang, K., 2023. Reservoir Inflow Forecasting in Hydropower Industry: A Generative Flow-Based Approach. *IEEE Transactions on Industrial Informatics* 19, 1196–1206. <https://doi.org/10.1109/TII.2022.3158927>

Zhou, Q., Chen, L., Singh, V.P., Zhou, J., Chen, X., Xiong, L., 2019. Rainfall-runoff simulation in karst dominated areas based on a coupled conceptual hydrological model. *Journal of Hydrology* 573, 524–533. <https://doi.org/10.1016/j.jhydrol.2019.03.099>

Appendix A

Temporal Hierarchical Reconciliation for Consistent Water Resources Forecasting Across Multiple Timescales: An Application to Precipitation Forecasting (Chapter 2)

This Appendix includes supporting information related to Chapter 2. More info regarding the reconciliation methods implemented is provided. In A1, the selected basins are listed, and the statistical parameters of the precipitation data (in the test set) are provided. Next, A2 provides details on the models' performance in the training set. Afterwards, A3 includes additional information related to temporal hierarchical reconciliation. Finally, in A4, the impact of the hierarchy's structure on the forecast results is clarified.

A1 Selected Basins' ID and Statistical Characteristics

The basins listed below in Table A 1 were selected for generating the forecasts utilized in this work.

Table A 1. Selected basins and their associated IDs.

CANOPEX ID	Basin
33	MUSKWA RIVER NEAR FORT NELSON
43	SIKANNI CHIEF RIVER NEAR FORT NELSON
79	NASS RIVER ABOVE SHUMAL CREEK
82	KISKATINAW RIVER NEAR FARMINGTON
86	LESSER SLAVE RIVER AT SLAVE LAKE
105	STUART RIVER NEAR FORT ST. JAMES
114	STELLAKO RIVER AT GLENANNAN
116	BEAVER RIVER NEAR DORINTOSH
132	BLACKMUD CREEK NEAR ELLERSLIE
180	ASSINIBOINE RIVER AT STURGIS
200	SEYMOUR RIVER NEAR SEYMOUR ARM
208	KICKING HORSE RIVER AT GOLDEN
212	ADAMS RIVER NEAR SQUILAX
214	CHILKO RIVER AT OUTLET OF CHILKO LAKE
216	EAGLE RIVER NEAR MALAKWA
217	TASEKO RIVER AT OUTLET OF TASEKO LAKES
226	MANIGOTAGAN RIVER NEAR MANIGOTAGAN

229	TURTLE RIVER NEAR LAURIER
236	ROLLING RIVER NEAR ERICKSON
242	LILLOOET RIVER NEAR PEMBERTON
245	SANDY CREEK NEAR CARON
268	SPIUS CREEK NEAR CANFORD
290	UPPER HUMBER RIVER NEAR REIDVILLE
291	LIGHTNING CREEK NEAR CARNDUFF
292	GRAHAM CREEK NEAR MELITA
295	SAGE CREEK AT Q RANCH NEAR WILDHORSE
296	ENGLISH RIVER AT UMFREVILLE
298	GRANBY RIVER AT GRAND FORKS
301	FLATHEAD RIVER AT FLATHEAD
303	ROCK CREEK BELOW HORSE CREEK NEAR INTERNATIONAL BOUNDARY
328	GANDER RIVER AT BIG CHUTE
347	UPSALQUITCH RIVER AT UPSALQUITCH
362	MATAWIN (RIVIERE) A SAINT-MICHEL-DES-SAINTS
366	SOUTHWEST MIRAMICHI RIVER AT BLACKVILLE
369	CANAAN RIVER AT EAST CANAAN
377	ST. MARYS RIVER AT STILLWATER
378	LAHAVE RIVER AT WEST NORTHFIELD
393	GRASS RIVER ABOVE STANDING STONE FALLS
396	BULKLEY RIVER NEAR HOUSTON
417	BARRIERE RIVER AT THE MOUTH
418	SHELL RIVER NEAR INGLIS
420	LITTLE SASKATCHEWAN RIVER NEAR MINNEDOSA
422	LARDEAU RIVER AT MARBLEHEAD
428	SLOCAN RIVER NEAR CRESCENT VALLEY
432	WHITEMOUTH RIVER NEAR WHITEMOUTH
454	WAPITI RIVER NEAR GRANDE PRAIRIE
461	QUESNEL RIVER AT LIKELY
462	RED DEER RIVER NEAR ERWOOD
469	CHILKO RIVER NEAR REDSTONE

470	COLUMBIA RIVER AT NICHOLSON
472	BROKENHEAD RIVER NEAR BEAUSEJOUR
482	SOUTH THOMPSON RIVER AT CHASE
484	ELK RIVER AT FERNIE
490	BULKLEY RIVER AT QUICK
494	QUESNEL RIVER NEAR QUESNEL
496	CLEARWATER RIVER NEAR CLEARWATER STATION
502	COLUMBIA RIVER AT DONALD
504	SMOKY RIVER AT WATINO
507	CHURCHILL RIVER NEAR PATUANAK
512	SKEENA RIVER AT USK
515	LIARD RIVER AT LOWER CROSSING
552	DORE RIVER NEAR MCBRIDE
578	BOW RIVER AT LAKE LOUISE
586	KOOTENAY RIVER AT KOOTENAY CROSSING
588	OCHRE RIVER AT OCHRE RIVER
602	KUSKANAX CREEK NEAR NAKUSP
605	TROUT CREEK NEAR GRANUM
606	BRIDGE CREEK AT GULL LAKE
608	KASLO RIVER BELOW KEMP CREEK
615	CROWSNEST RIVER AT FRANK
622	MANYBERRIES CREEK AT BRODIN'S FARM
623	DENNIEL CREEK NEAR VAL MARIE
624	CHEHALIS RIVER NEAR HARRISON MILLS
630	ENGLISHMAN RIVER NEAR PARKSVILLE
634	BIG SHEEP CREEK NEAR ROSSLAND
639	CHEMAINUS RIVER NEAR WESTHOLME
659	MONTMORENCY (RIVIERE) A 0
664	NORTHEAST MARGAREE RIVER AT MARGAREE VALLEY
667	SOUTHWEST MARGAREE RIVER NEAR UPPER MARGAREE
674	NORTH MAGNETAWAN RIVER NEAR BURK'S FALLS
676	BEAR BROOK NEAR BOURGET
688	ROSEWAY RIVER AT LOWER OHIO

695	MIDDLE THAMES RIVER AT THAMESFORD
698	KETTLE CREEK AT ST. THOMAS

In Table A 2, the statistical parameters mean, median, standard deviation, 95th percentile, and 5th percentile of the precipitation data (associated with the test set; 2005-2010) are based on average values considering all 84 basins, for each of the six timescales.

Table A 2. The test set statistical properties at different timescales.

Timescale	Statistical parameter (mm)				
	Mean	Median	Standard Deviation	95 th percentile	5 th percentile
Monthly	63.81	56.87	35.91	129.97	18.07
2-Monthly	127.63	120.22	55.74	222.85	54.47
Quarterly	191.44	182.16	75.24	310.57	90.26
4-Monthly	255.26	247.93	88.75	389.47	138.84
Bi-annual	382.89	391.24	86.93	490.31	256.05
Annual	765.78	783.71	131.18	903.12	599.92

A2 Training Error Evaluation

The Nash-Sutcliffe efficiency (NSE) values associated with the test set are presented in Table A 3. The results indicate that by aggregating the time-series from monthly up to 4-monthly scales, the accuracy of the models' increases, while a decrease in accuracy is observed from 4-monthly to annual scales. The same pattern was observed for the test set (see Table 2-2 in the main text). Except for monthly and annual timescales, where exponential time-series smoothing (ETS) and seasonal autoregressive integrated moving average (SARIMA) showed the best performance, respectively, artificial neural network (ANN) outperformed ETS and SARIMA. It is noted that the results for the training set (unlike the test set) are associated with 1-step ahead forecasting.

Table A 3. NSE values associated with the training set for ETS, ANN, and SARIMA.

Timescale	NSE		
	ETS	ANN	SARIMA
Monthly	0.29	0.25	0.22
2-Monthly	0.38	0.39	0.36
Quarterly	0.41	0.43	0.39
4-Monthly	0.42	0.44	0.41
Biannual	0.18	0.23	0.19
Annual	0.07	0.07	0.22

A3 Temporal Hierarchical Reconciliation

The annual sum of one-year ahead precipitation forecasts is shown in Table A 4 for a single basin (CANOPEX id: 33, see Table A 1). The forecasts were obtained using the ETS method. As can be seen, with an annual forecast horizon, the reconciliation methods resulted in consistent forecasts for all timescales. For this example, the monthly and annual forecasts had the highest deviation from the target (observed) and thus, benefitted most from temporal hierarchical reconciliation. Based on its definition, the BU method is equal to the original forecast (ORF) at the monthly scale (see Section 2.5 in the main text).

Table A 4. Annual sum of target values and ORF along with reconciled forecasts generated by the reconciliation methods: structural weighted least squares (WLS_S), variance weighted least squares (WLS_V), ordinary least squares (OLS), and the bottom-up (BU) method.

Type	Annual sum (mm, for one year)					
	Monthly (12-steps ahead)	2-Monthly (6-steps ahead)	Quarterly (4-steps ahead)	4-Monthly (3-steps ahead)	Bi-annual (2-steps ahead)	Annual (1-step ahead)
Target value	573.77	573.77	573.77	573.77	573.77	573.77
ORF	529.64	571.21	572.04	572.75	572.4	590.86
WLS _S	568.15	568.15	568.15	568.15	568.15	568.15
WLS _V	571.75	571.75	571.75	571.75	571.75	571.75
OLS	578.71	578.71	578.71	578.71	578.71	578.71
BU	529.64	529.64	529.64	529.64	529.64	529.64

Average NSE values of the reconciled forecasts are listed in Table A5. At each timescale, the results are associated with ORFs having an NSE greater than 0 (filtered results). The results show that for ETS at timescales up to 4-monthly, the BU method outperformed the temporal hierarchical reconciliation methods. For ANN and SARIMA, it is observed that utilizing the BU method resulted in the most erroneous forecasts while the hierarchical methods improved accuracy.

Table A 5. Average NSE values of the reconciliation methods obtained by each model (ETS, ANN, and SARIMA). Based on the definition of the BU method, the monthly scale values listed for ORF and BU are equal. At each timescale, the results are associated with ORFs having a NSE greater than 0.

Method	Monthly	2-Monthly	Quarterly	4-Monthly	Bi-annual	Annual
ETS						
WLS _S	0.297	0.393	0.433	0.483	0.318	0.078
WLS _V	0.292	0.386	0.427	0.480	0.322	0.098
OLS	0.291	0.387	0.423	0.480	0.323	0.091
BU	0.305	0.403	0.444	0.482	0.309	0.091
ANN						
WLS _S	0.273	0.382	0.384	0.477	0.250	0.069
WLS _V	0.274	0.382	0.385	0.478	0.251	0.066
OLS	0.273	0.386	0.389	0.488	0.258	0.070
BU	0.246	0.330	0.320	0.409	0.165	-0.915
SARIMA						
WLS _S	0.275	0.377	0.410	0.506	0.277	0.057
WLS _V	0.276	0.379	0.410	0.508	0.276	0.061
OLS	0.276	0.382	0.412	0.516	0.286	0.088
BU	0.212	0.271	0.299	0.374	0.167	0.070

A3.1 Analyzing the Formulation of the Temporal Hierarchical Reconciliation Methods

Based on the generalized least squares method (Eq 2-10 in the main text) (Athanasopoulos et al., 2017):

$$\beta_i^{\text{GLS}}(h) = (\mathbf{S}'\Sigma_h^{-1}\mathbf{S})^{-1}\mathbf{S}'\Sigma_h^{-1}\hat{\mathbf{y}}_i(h) \quad \text{Eq A 1}$$

where Σ_h is the covariance matrix, and the variable which controls the impact of WLS_S, WLS_V, and OLS on the resulting forecasts. The three temporal hierarchical reconciliation methods adopted in this study (WLS_S, WLS_V, and OLS) only differ due to their definition of Σ_h . For WLS_S and OLS, Σ_h is not dependent on the forecast models. Next, the matrix Ψ can be defined as:

$$\Psi = S(S'\Sigma_h^{-1}S)^{-1}S'\Sigma_h^{-1} \quad \text{Eq A 2}$$

where for WLS_S and OLS, Ψ is independent of the adopted model. The Ψ matrix is the weights matrix in a linear regression model, which designates how much each timescale is impacted by other timescales in the temporal hierarchy. Values of Ψ for WLS_S and OLS as well as the absolute difference between the two, are shown in Fig A 1. Each row/column in Fig A 1 corresponds to the position of the forecast in the hierarchy (e.g., row/column 1 is associated with the annual scale while rows/columns 17-28 are associated with the monthly scale; see Figure 2-2 in the main text). For both methods, monthly forecasts (rows/columns 17-28) are only slightly impacted by more coarsely aggregated timescales and tend to impact other timescales significantly. The results indicate that by using OLS, more weight is given to more coarsely aggregated series for reconciliation compared to WLS_S. This explains why, for SARIMA and ANN, in general, OLS resulted in more accurate forecasts, as there was a significant gap between the accuracy of the ORFs at the monthly scale and the more coarsely aggregated timescales (e.g., 2-monthly, and 4-monthly), which can be clearly seen in Table A 5. For ETS, however, WLS_S performed superior to OLS, as the ORFs at the monthly scale were considered accurate, and the gap between the accuracy of coarser timescales and the monthly series was not as significant as it was for ANN and SARIMA.

A3.2 Comparison of Temporal Hierarchical Reconciliation Methods

The results in the main text showed that the temporal hierarchical reconciliation methods (WLS_S, WLS_V, and OLS), coupled with the different forecasting models, resulted in dissimilar forecast accuracy. For each forecast model type (ETS, ANN, SARIMA), the basin-averaged nRMSE values associated with the filtered results at each timescale (NSE greater than 0) are included in Fig A 2. Recall that the BU method at the monthly scale is equivalent to ORF, and lower nRMSE values indicate improved performance (compared to competing methods). The results indicate that the WLS_S, WLS_V, and OLS temporal hierarchical reconciliation methods do not differ much from the BU method for the ETS forecasts. For the ETS forecasts, the BU method followed by the WLS_S method slightly outperformed WLS_V and OLS. For ANN and SARIMA forecasts, BU had the worst performance,

whereas the OLS method showed better performance than WLS_S and WLS_V . Although it was observed that temporal hierarchical reconciliation methods have a somewhat similar impact on "inaccurate" forecasts as on "accurate" forecasts, the accuracy of the reconciled forecasts is more heavily influenced by the ORF accuracy than it is by the type of temporal hierarchical reconciliation method. A figure like Figure 2-8 in the main text is provided here, where the y-axis is different for each method (subfigure; Fig A 3). The figure confirms that the BU method generally shows superior performance to the temporal hierarchical reconciliation methods when combined with ETS, especially for less accurate forecasts at the monthly level. The results indicate that a single best method could not be selected for temporal hierarchical reconciliation, and the accuracy of the methods is case and timescale dependent, suggesting that forecasts should be tailored specifically for a given basin.

A4 Temporal Hierarchy's Structure and its Impact on Reconciliation

The average NSE values (filtered based on the ORFs' NSE) associated with the new setup (considering only three timescales, namely monthly, 2-monthly, and 4-monthly) are listed in Table A 6. Comparing values in Table A 6 with values listed in Table A 3, it is understood that, generally, for ETS, the accuracy has slightly increased, while for ANN and SARIMA, the performance has slightly decreased.

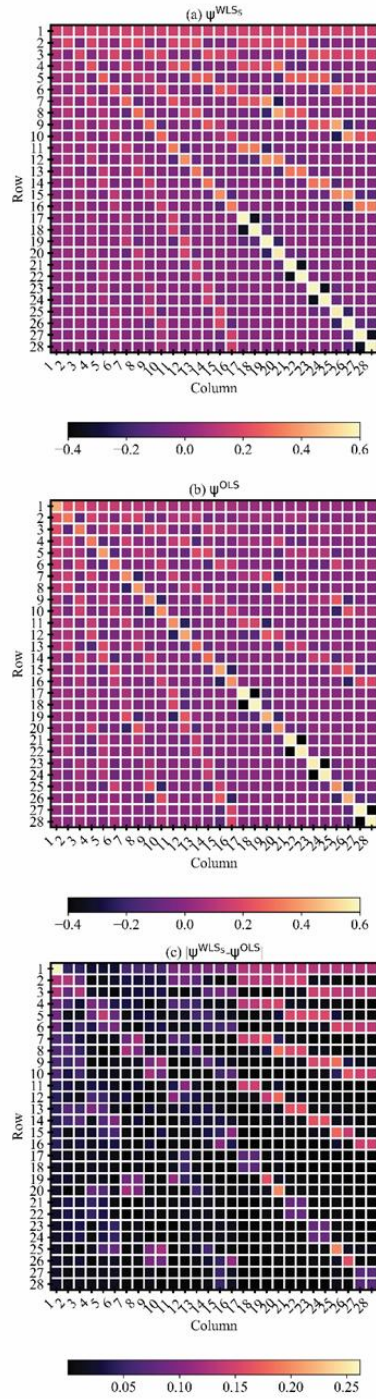


Fig A 1. Weight matrices: (a) Ψ matrix for WLS (Ψ^{WLS}) (b) Ψ matrix for OLS (Ψ^{OLS}), and (c) the absolute difference between Ψ^{WLS} and Ψ^{OLS} .

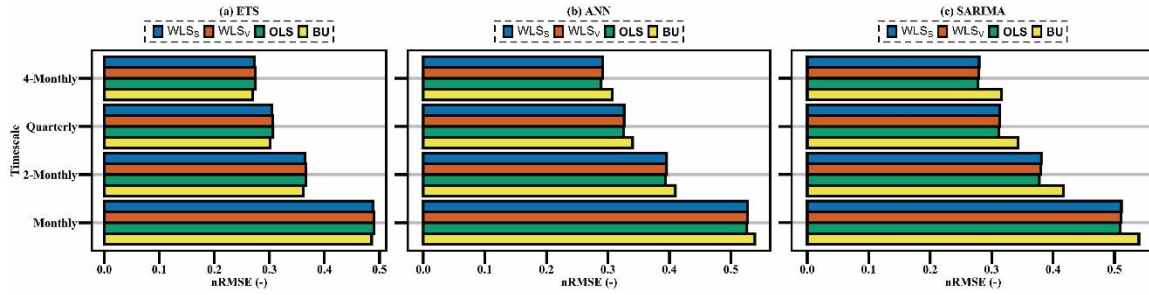


Fig A 2. Comparison of temporal hierarchical reconciliation methods based on basin-averaged nRMSE values for (a) ETS, (b) ANN, and (c) SARIMA. For comparison, only forecasts with NSE greater than 0 were considered at each timescale.

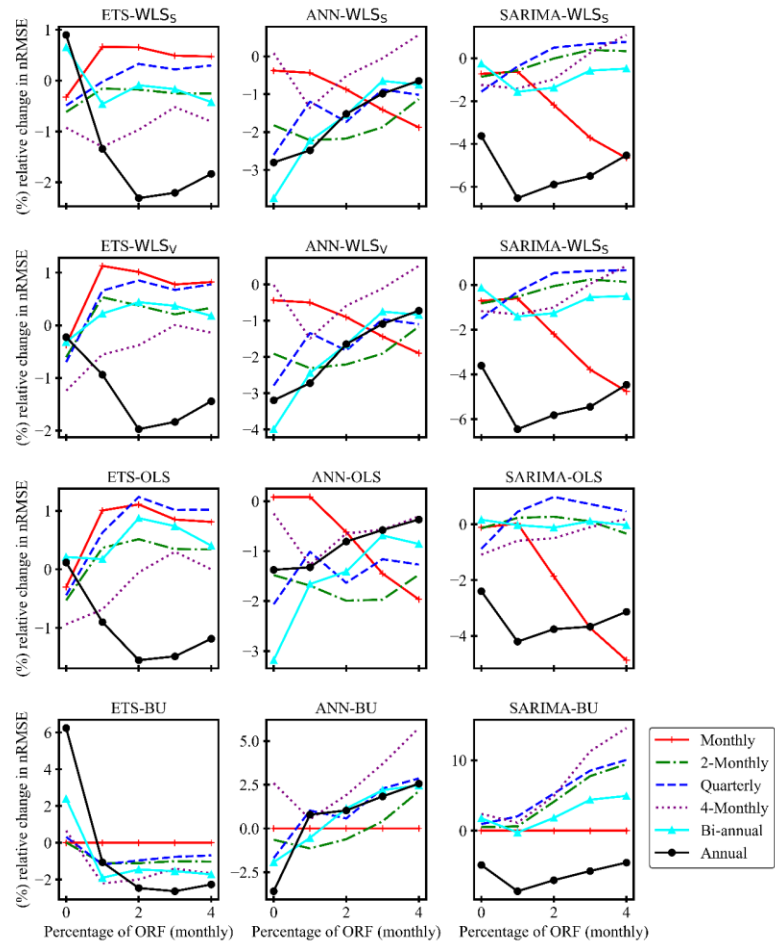


Fig A 3. The relative $([\text{reconciled-ORF}]/\text{ORF} \times 100\%)$ nRMSE change of the temporal hierarchical reconciliation methods compared to ORFs at different timescales. The results are associated with ORFs having an NSE value greater than zero at the monthly scale. The x-axis shows the accumulated percentage of sorted

(based on nRMSE) monthly forecasts. Note: that BU is equal to ORF at the monthly scale resulting in no relative change.

Table A 6. Average NSE values obtained through new hierarchy setup, obtained by each model (ETS, ANN, and SARIMA). At each timescale, the results are associated with ORFs having an NSE greater than 0.

Method	Monthly	2-Monthly	4-Monthly
ETS			
WLS _S	0.299	0.396	0.484
WLS _V	0.293	0.386	0.480
OLS	0.296	0.391	0.483
Method	ANN		
WLS _S	0.268	0.377	0.476
WLS _V	0.269	0.379	0.479
OLS	0.269	0.381	0.487
Method	SARIMA		
WLS _S	0.269	0.367	0.497
WLS _V	0.272	0.372	0.505
OLS	0.274	0.380	0.518

Appendix B

Generative deep learning for probabilistic streamflow forecasting: conditional variational auto-encoder (Chapter 3)

This Appendix includes supporting information related to Chapter 3. In B1, Quantile-based conditional variational auto-encoder is explained. In B2, Results related to hyperparameter selection are provided. In B3 and B4, deterministic and probabilistic evaluations results are presented. It is noted that supplementary data to the associated chapter can be found online at <https://doi.org/10.1016/j.jhydrol.2023.130498>. The forecast results can be found online at: <https://zenodo.org/doi/10.5281/zenodo.10223595>.

B1 Quantile-Based Conditional Variational Auto-Encoder (QCVAE)

Unlike the CVAE network, where the ELBO function is maximized, the following loss function is minimized:

$$L_{\Phi, \phi}^t = D_{KL}(p_{\Phi}(\mathbf{z}^t | \mathbf{x}^t) | p(\mathbf{z}^t)) + \mathbb{E}_{p_{\Phi}} \left(\sum L_{q_i}^t \right) \quad \text{Eq B 1}$$

where $q_i \in \{0.05, 0.5, 0.95\}$. While the variational term (D_{KL}) is still present in the loss function, the log-likelihood term is substituted with the pinball loss $L_{q_i}^t$.

B2 Hyperparameter Selection

Hyperparameter selection for CVAE^Q and QCVAE^Q are shown in Fig B 1 and Fig B 2, respectively.

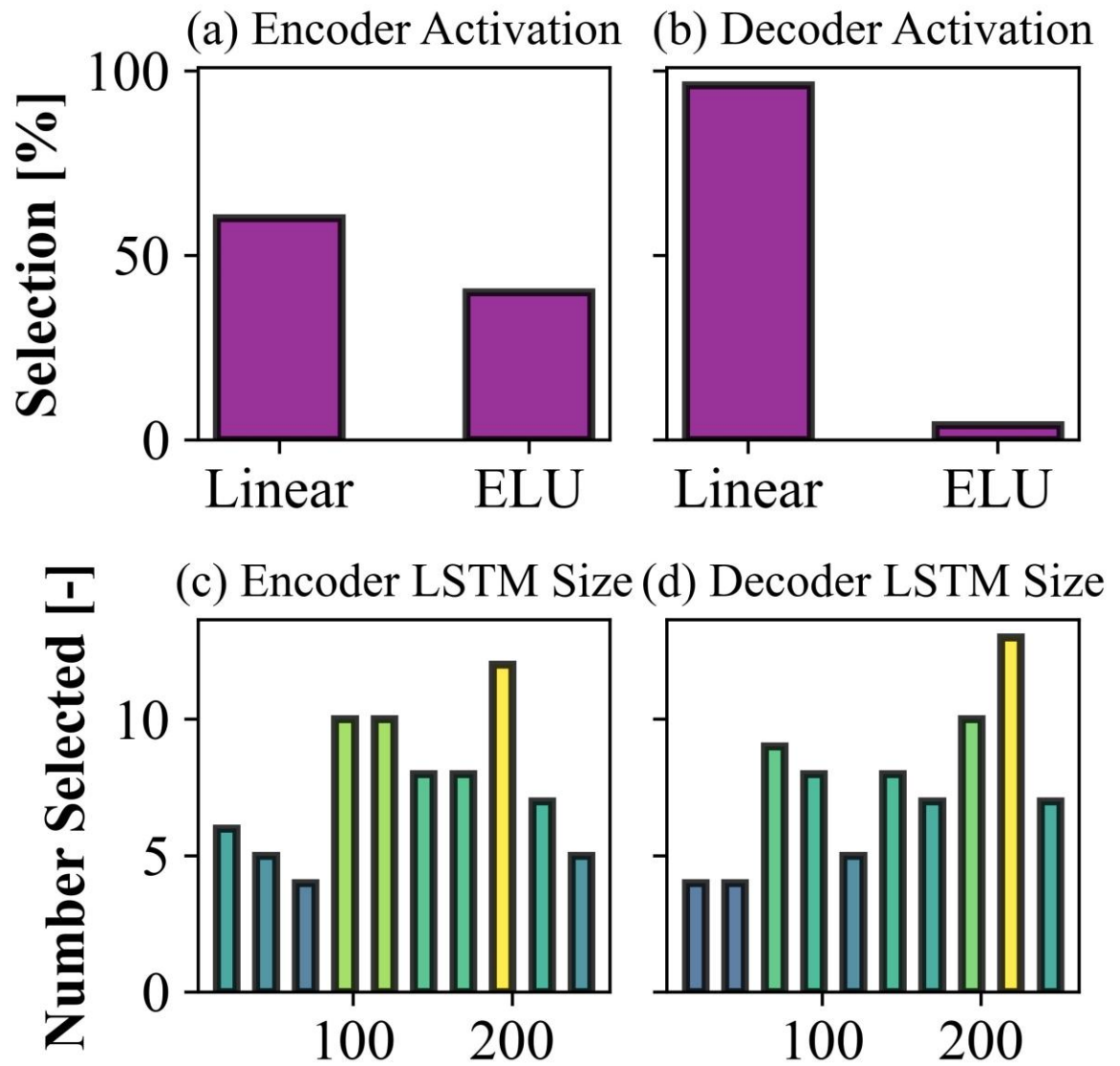


Fig B 1. Hyperparameter selection (CVAE^Q) analysis. The first column is associated with the encoder, while the second is related to the decoder structure.

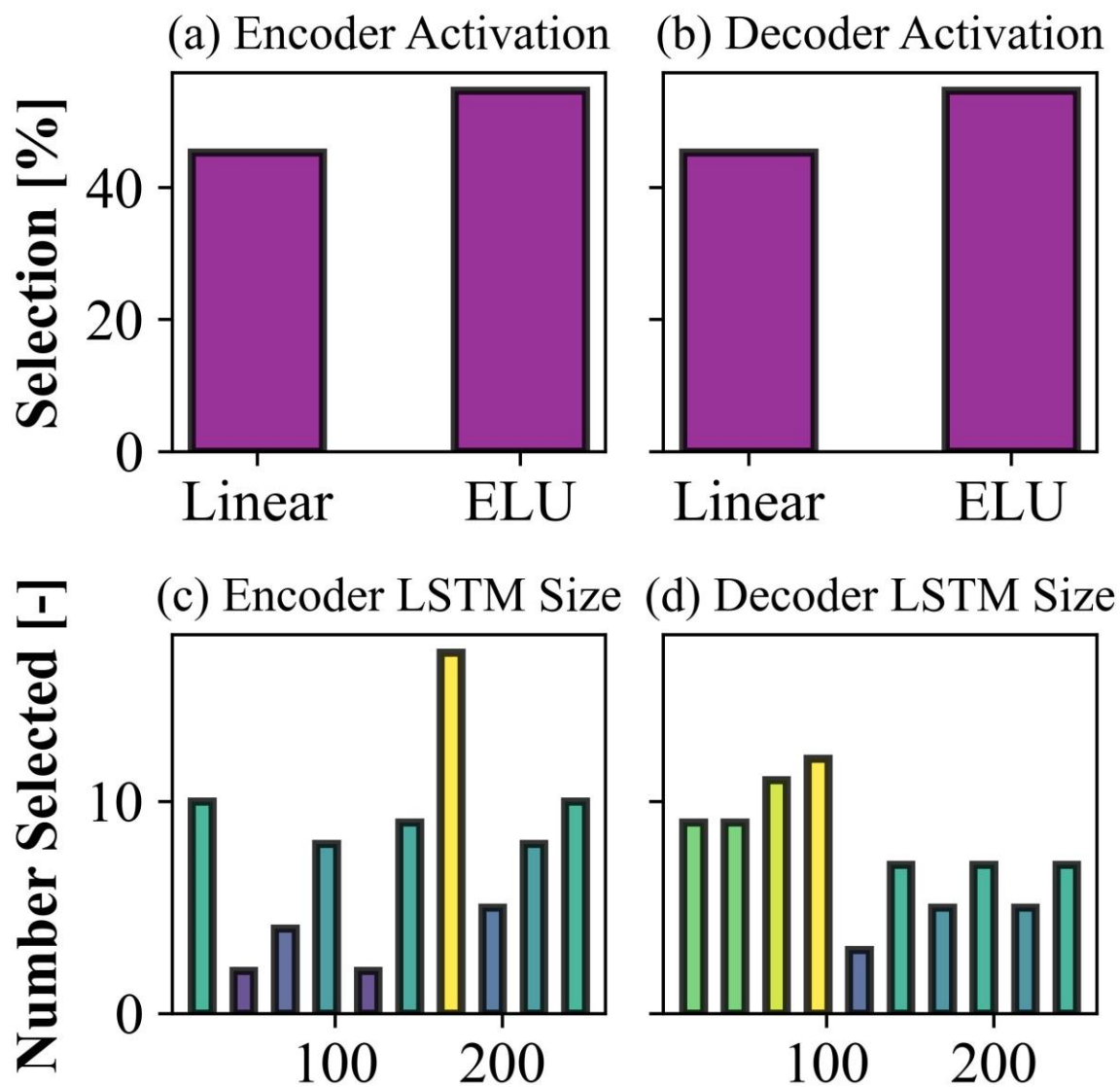


Fig B 2. Hyperparameter selection (QCVAE^Q) analysis. The first column is associated with the encoder, while the second is related to the decoder structure.

B3 Deterministic Evaluation

In Table B 1, stepwise comparison of the KGE scores of the forecasts relative to the simulations is provided.

Table B 1. Stepwise comparison of the KGE scores of the forecasts relative to the simulations $((KGE_{\text{forecast model}}^{\text{day}} - KGE_{\text{simulation}})/KGE_{\text{simulation}} \times 100\%)$. KGE scores were averaged across all 75 basins.

Forecast Model	KGE change [%]						
	day 1	day 2	day 3	day 4	day 5	day 6	day 7
ED ^{NQ}	1.62	-2.11	-7.42	-10.92	-13.91	-16.24	-18.39
ED ^Q	24.32	16.97	9.02	3.10	-1.38	-4.95	-8.04
CVAE ^{NQ}	-8.40	-10.31	-13.64	-16.78	-18.65	-20.89	-22.48
CVAE ^Q	19.68	13.47	6.83	1.33	-3.03	-6.56	-9.58
QCVAE ^{NQ}	2.00	-1.82	-7.27	-10.96	-13.87	-16.11	-17.94
QCVAE ^Q	23.98	16.48	8.58	2.85	-1.32	-4.86	-7.56

B4 Probabilistic Evaluation

Fig B 3 displays probabilistic forecasts for a randomly selected basin (specifically, basin 01BJ003) associated with three different lead times (LT = 1, 3, and 7). In contrast to the other two types of forecast models, there is little discernible difference between the PICP values produced by the CVAE^Q and QCVAE^Q models. In other words, the CVAE model appears reliable even when the forecast accuracy is relatively low. Consistent with the results presented in Figure 3-3 and Table 3-2, the ED and QCVAE models demonstrate better forecast accuracy compared to the CVAE model.

The impact of different features of the input variables on the performance of ED^Q is illustrated in Fig B 4. To check whether the feature has an impact on the results or not, a regression line was fitted, and the significance of its coefficient was analyzed.

The relationship between KGE and PICP (i.e., the deterministic and probabilistic metrics) for ED^Q is depicted in Fig B 5. Unlike CVAE^Q forecasts, no significant relationship was observed between the two types of metrics.

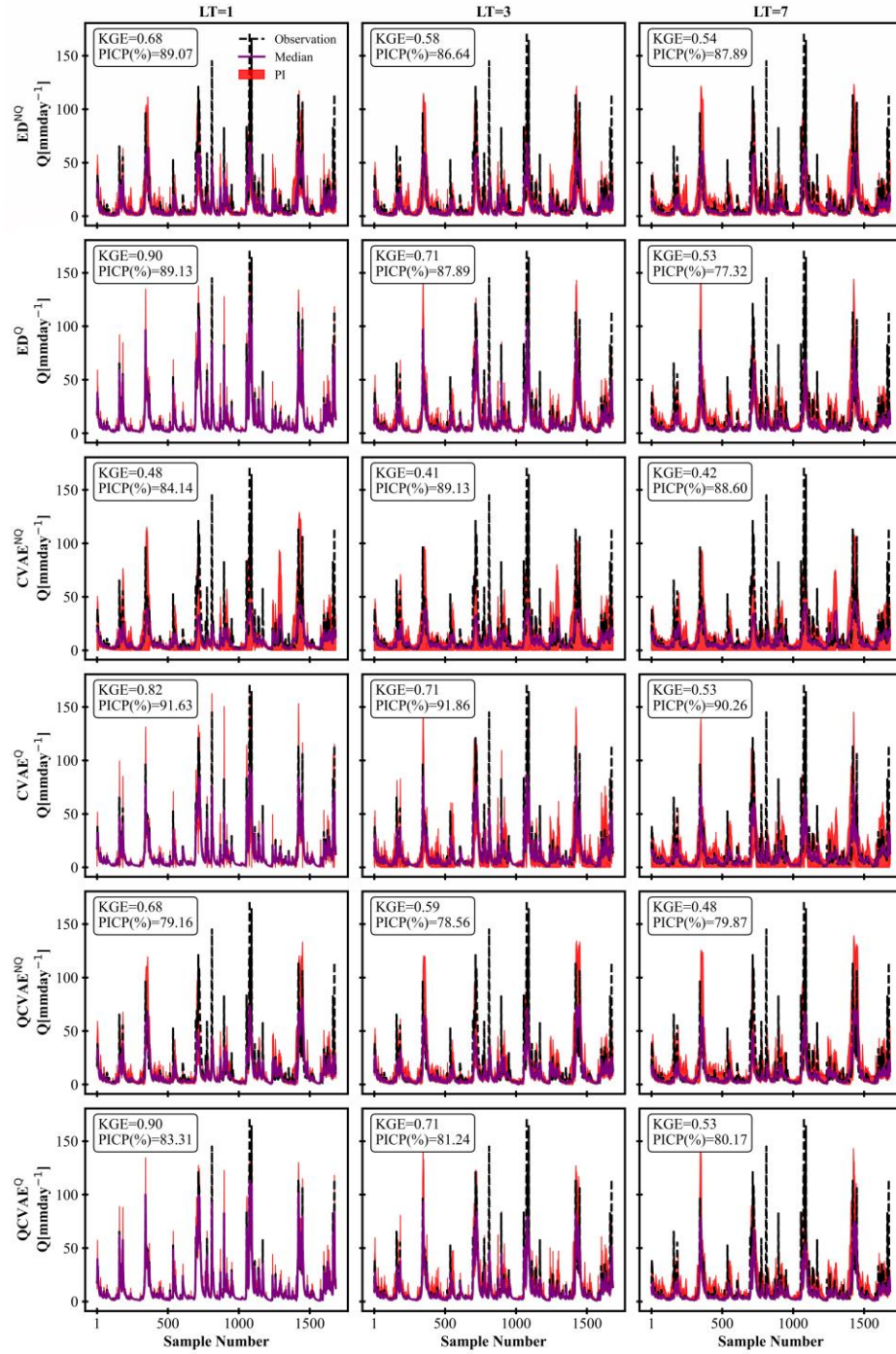


Fig B 3. Prediction interval (PI) estimates associated with LT=1 (column 1), LT=3 (column 2), and LT=7 (column 3) for a randomly selected basin.

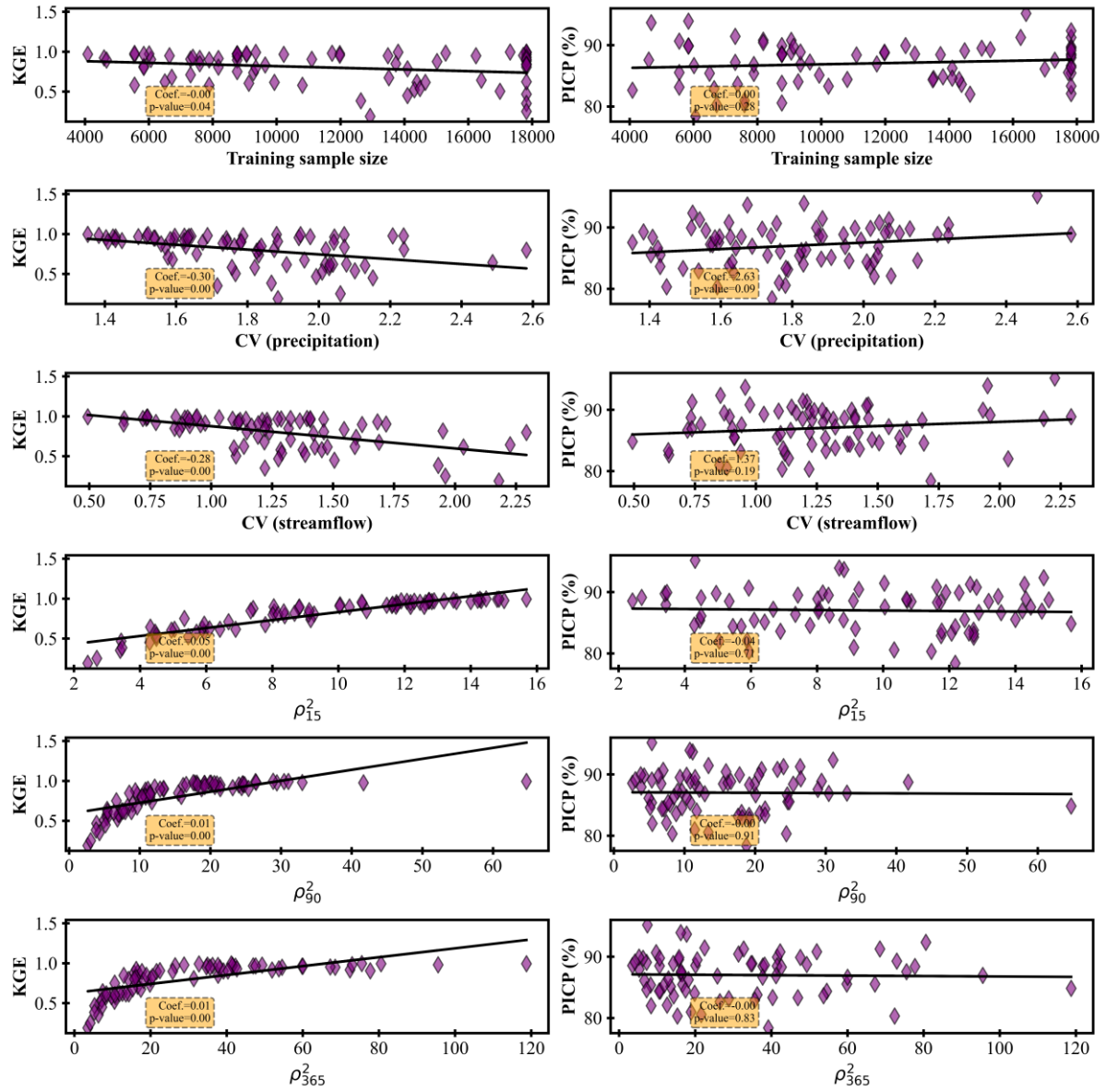


Fig B 4. Impact of different features of the input variables on the performance of ED^Q. KGE and PICP values are averaged across forecast lead times of one-seven days ahead. The fitted lines are based on linear regression where their associated coefficient (slope coefficient) and the associated p-values are shown in the boxes.

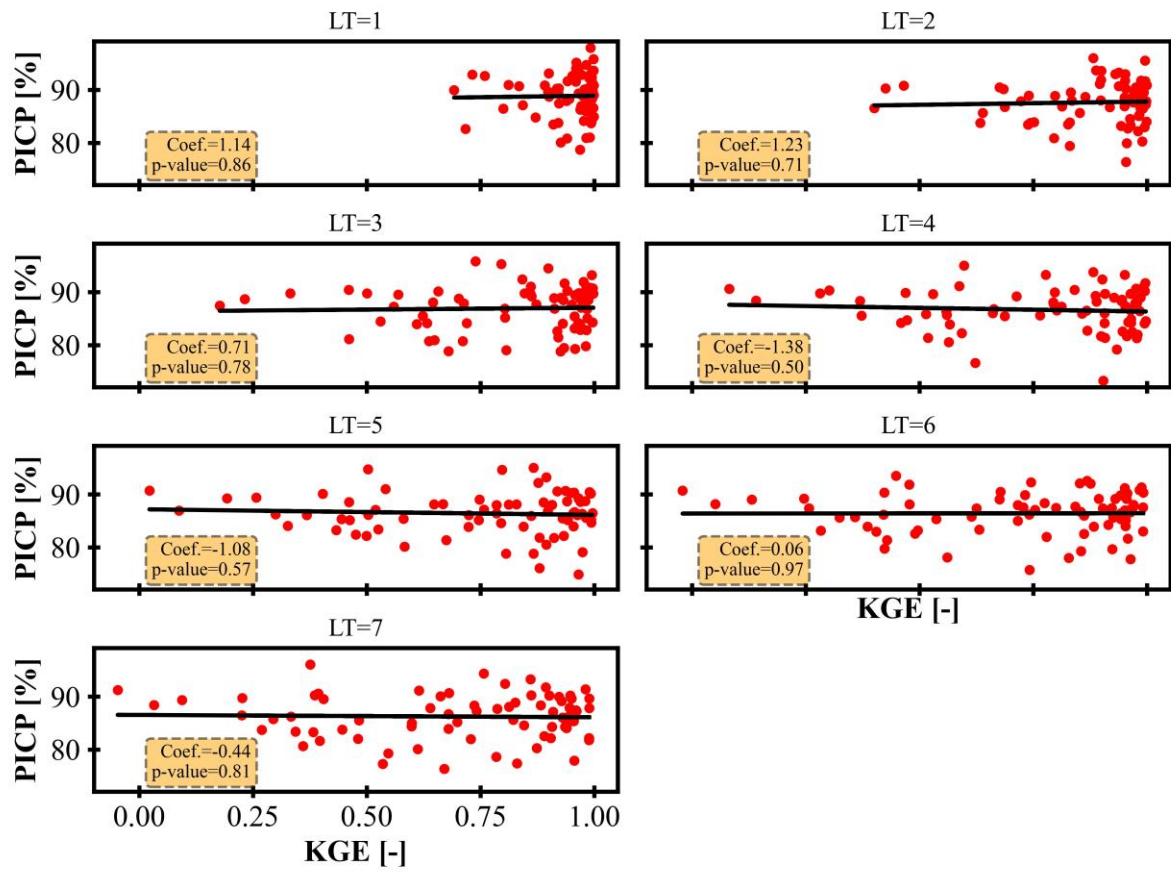


Fig B 5. KGE-PICP for different LTs associated with ED^Q . The slope coefficients and their associated p-values are shown in the box.

Appendix C

Hierarchical Deep Learning for Consistent Multi-Timescale Hydrological Forecasting (Chapter 4)

This Appendix includes supporting information related to Chapter 4. More information regarding the deep learning (DL) models is provided. In C1, the selected catchments are listed, and their distribution across the continental U.S. (CONUS) is shown. In C2, the formulation of the LSTM model and error metrics are provided. Afterward, C3 includes additional information about the models' multi-step ahead forecast performance. Next, C4 details the models' simulation performance using Daymet and ERA5-Land data. Finally, in C5, the performance of the weekly forecasts' prior to temporal hierarchical reconciliation (THR) is shown.

C1 Data

Selected Catchments' ID and Outlet Dispersion

The catchments listed below were selected for generating the forecasts in this work.

Table C 1. Selected basins and their associated camels IDs.

Gauge id	Gauge name
camels_01022500	Narraguagus River at Cherryfield, Maine
camels_01031500	Piscataquis River near Dover-Foxcroft, Maine
camels_01047000	Carrabassett River near North Anson, Maine
camels_01052500	Diamond River near Wentworth Location, NH
camels_01054200	Wild River at Gilead, Maine
camels_01055000	Swift River near Roxbury, Maine
camels_01057000	Little Androscoggin River near South Paris, Maine
camels_01078000	SMITH RIVER NEAR BRISTOL, NH
camels_01134500	MOOSE RIVER AT VICTORY, VT
camels_01137500	AMMONOOSUC RIVER AT BETHLEHEM JUNCTION, NH
camels_01139000	WELLS RIVER AT WELLS RIVER, VT
camels_01144000	WHITE RIVER AT WEST HARTFORD, VT
camels_01169000	NORTH RIVER AT SHATTUCKVILLE, MA
camels_01170100	GREEN RIVER NEAR COLRAIN, MA
camels_01181000	WEST BRANCH WESTFIELD RIVER AT HUNTINGTON, MA

camels_01333000	GREEN RIVER AT WILLIAMSTOWN, MA
camels_01350000	SCHOHARIE CREEK AT PRATTSVILLE NY
camels_01365000	RONDOUT CREEK NEAR LOWES CORNERS NY
camels_01413500	EAST BR DELAWARE R AT MARGARETVILLE NY
camels_01423000	WEST BRANCH DELAWARE RIVER AT WALTON NY
camels_01435000	NEVERSINK RIVER NEAR CLARYVILLE NY
camels_01440000	FLAT BROOK NEAR FLATBROOKVILLE NJ
camels_01440400	Brodhead Creek near Analomink, PA
camels_01451800	Jordan Creek near Schnecksville, PA
camels_01485500	NASSAWANGO CREEK NEAR SNOW HILL, MD
camels_01487000	NANTICOKE RIVER NEAR BRIDGEVILLE, DE
camels_01491000	CHOPTANK RIVER NEAR GREENSBORO, MD
camels_01518862	Cowanesque River at Westfield, PA
camels_01532000	Towanda Creek near Monroeton, PA
camels_01539000	Fishing Creek near Bloomsburg, PA
camels_01543000	Driftwood Br Sinnemahoning Cr at Sterling Run, PA
camels_01543500	Sinnemahoning Creek at Sinnemahoning, PA
camels_01544500	Kettle Creek at Cross Fork, PA
camels_01545600	Young Womans Creek near Renovo, PA
camels_01547700	Marsh Creek at Blanchard, PA
camels_01548500	Pine Creek at Cedar Run, PA
camels_01550000	Lycoming Creek near Trout Run, PA
camels_01552000	Loyalsock Creek at Loyalsockville, PA
camels_01557500	Bald Eagle Creek at Tyrone, PA
camels_01568000	Sherman Creek at Shermans Dale, PA
camels_01580000	DEER CREEK AT ROCKS, MD
camels_01583500	WESTERN RUN AT WESTERN RUN, MD
camels_01596500	SAVAGE RIVER NEAR BARTON, MD
camels_01605500	SOUTH BRANCH POTOMAC RIVER AT FRANKLIN, WV
camels_01632000	N F SHENANDOAH RIVER AT COOTES STORE, VA
camels_01632900	SMITH CREEK NEAR NEW MARKET, VA
camels_01634500	CEDAR CREEK NEAR WINCHESTER, VA
camels_01638480	CATOCTIN CREEK AT TAYLORSTOWN, VA

camels_01639500	BIG PIPE CREEK AT BRUCEVILLE, MD
camels_01644000	GOOSE CREEK NEAR LEESBURG, VA
camels_01664000	RAPPAHANNOCK RIVER AT REMINGTON, VA
camels_01666500	ROBINSON RIVER NEAR LOCUST DALE, VA
camels_01667500	RAPIDAN RIVER NEAR CULPEPER, VA
camels_01669520	DRAGON SWAMP AT MASCOT, VA
camels_02011400	JACKSON RIVER NEAR BACOVA, VA
camels_02011460	BACK CREEK NEAR SUNRISE, VA
camels_02013000	DUNLAP CREEK NEAR COVINGTON, VA
camels_02014000	POTTS CREEK NEAR COVINGTON, VA
camels_02015700	BULLPASTURE RIVER AT WILLIAMSVILLE, VA
camels_02016000	COWPASTURE RIVER NEAR CLIFTON FORGE, VA
camels_02017500	JOHNS CREEK AT NEW CASTLE, VA
camels_02018000	CRAIG CREEK AT PARR, VA
camels_02027000	TYE RIVER NEAR LOVINGSTON, VA
camels_02027500	PINEY RIVER AT PINEY RIVER, VA
camels_02028500	ROCKFISH RIVER NEAR GREENFIELD, VA
camels_02046000	STONY CREEK NEAR DINWIDDIE, VA
camels_02051000	NORTH MEHERRIN RIVER NEAR LUNENBURG, VA
camels_02051500	MEHERRIN RIVER NEAR LAWRENCEVILLE, VA
camels_02053200	POTECASI CREEK NEAR UNION, NC
camels_02053800	S F ROANOKE RIVER NEAR SHAWSVILLE, VA
camels_02056900	BLACKWATER RIVER NEAR ROCKY MOUNT, VA
camels_02059500	GOOSE CREEK NEAR HUDDLESTON, VA
camels_02064000	FALLING RIVER NEAR NARUNA, VA
camels_02065500	CUB CREEK AT PHENIX, VA
camels_02069700	SOUTH MAYO RIVER NEAR NETTLERIDGE, VA
camels_02070000	NORTH MAYO RIVER NEAR SPENCER, VA
camels_02074500	SANDY RIVER NEAR DANVILLE, VA
camels_02077200	HYCO CREEK NEAR LEASBURG, NC
camels_02081500	TAR RIVER NEAR TAR RIVER, NC
camels_02082950	LITTLE FISHING CREEK NEAR WHITE OAK, NC
camels_02108000	NORTHEAST CAPE FEAR RIVER NEAR CHINQUAPIN, NC

camels_02111180	ELK CREEK AT ELKVILLE, NC
camels_02111500	REDDIES RIVER AT NORTH WILKESBORO, NC
camels_02112120	ROARING RIVER NEAR ROARING RIVER, NC
camels_02118500	HUNTING CREEK NEAR HARMONY, NC
camels_02128000	LITTLE RIVER NEAR STAR, NC
camels_02137727	CATAWBA R NR PLEASANT GARDENS, NC
camels_02143000	HENRY FORK NEAR HENRY RIVER, NC
camels_02149000	COVE CREEK NEAR LAKE LURE, NC
camels_02152100	FIRST BROAD RIVER NEAR CASAR, NC
camels_02177000	CHATTOOGA RIVER NEAR CLAYTON, GA
camels_02212600	FALLING CREEK NEAR JULIETTE, GA
camels_02216180	TURNPIKE CREEK NEAR MCRAE, GA
camels_02221525	MURDER CREEK BELOW EATONTON, GA
camels_02231000	ST. MARYS RIVER NR MACCLENNY, FLA.
camels_02245500	SOUTH FORK BLACK CREEK NR PENNEY FARMS, FLA.
camels_02246000	NORTH FORK BLACK CREEK NR MIDDLEBURG, FLA.
camels_02296500	CHARLIE CREEK NEAR GARDNER FL
camels_02297310	HORSE CREEK NEAR ARCADIA FL
camels_02298123	PRAIRIE CREEK NEAR FORT OGDEN FL
camels_02298608	MYAKKA RIVER AT MYAKKA CITY FL
camels_02299950	MANATEE RIVER NEAR MYAKKA HEAD FL
camels_02312200	LITTLE WITHLACOOCHIE RIVER AT RERDELL, FLA.
camels_02324400	FENHOLLOWAY RIVER NEAR FOLEY, FLA.
camels_02327100	SOPCHOPPY RIVER NR SOPCHOPPY, FLA.
camels_02349900	TURKEY CREEK AT BYROMVILLE, GA
camels_02361000	CHOCTAWHATCHEE RIVER NEAR NEWTON, AL.
camels_02369800	BLACKWATER RIVER NEAR BRADLEY AL
camels_02371500	CONECUH RIVER AT BRANTLEY AL
camels_02372250	PATSALIGA CREEK NEAR BRANTLEY AL
camels_02374500	MURDER CREEK NEAR EVERGREEN AL
camels_02408540	HATCHET CREEK BELOW ROCKFORD AL
camels_02422500	MULBERRY CREEK AT JONES AL
camels_02450250	SIPSEY FORK NEAR GRAYSON AL

camels_02464000	NORTH RIVER NEAR SAMANTHA AL
camels_02469800	SATILPA CREEK NEAR COFFEEVILLE AL
camels_02472000	LEAF RIVER NR COLLINS, MS
camels_02472500	BOUIE CREEK NR HATTIESBURG, MS
camels_02479155	CYPRESS CREEK NR JANICE, MS
camels_02479560	ESCATAWPA RIVER NEAR AGRICOLA MS
camels_02481000	BILOXI RIVER AT WORTHAM, MS
camels_02481510	WOLF RIVER NR LANDON, MS
camels_03010655	Oswayo Creek at Shinglehouse, PA
camels_03011800	Kinzua Creek near Guffey, PA
camels_03015500	Brokenstraw Creek at Youngsville, PA
camels_03021350	French Creek near Wattsburg, PA
camels_03028000	West Branch Clarion River at Wilcox, PA
camels_03049000	Buffalo Creek near Freeport, PA
camels_03069500	CHEAT RIVER NEAR PARSONS, WV
camels_03070500	BIG SANDY CREEK AT ROCKVILLE, WV
camels_03076600	BEAR CREEK AT FRIENDSVILLE, MD
camels_03078000	CASSELMAN RIVER AT GRANTSVILLE, MD
camels_03144000	Wakatomika Creek near Frazeyburg OH
camels_03165000	CHESTNUT CREEK AT GALAX, VA
camels_03170000	LITTLE RIVER AT GRAYSONTOWN, VA
camels_03173000	WALKER CREEK AT BANE, VA
camels_03180500	GREENBRIER RIVER AT DURBIN, WV
camels_03182500	GREENBRIER RIVER AT BUCKEYE, WV
camels_03186500	WILLIAMS RIVER AT DYER, WV
camels_03238500	White Oak Creek near Georgetown OH
camels_03241500	Massies Creek at Wilberforce OH
camels_03280700	CUTSHIN CREEK AT WOOTON, KY
camels_03281500	SOUTH FORK KENTUCKY RIVER AT BOONEVILLE, KY
camels_03285000	DIX RIVER NEAR DANVILLE, KY
camels_03340800	BIG RACCOON CREEK NEAR FINCASTLE IND
camels_03346000	NORTH FORK EMBARRAS RIVER NEAR OBLONG, IL
camels_03364500	CLIFTY CREEK AT HARTSVILLE, IN

camels_03366500	MUSCATATUCK RIVER NEAR DEPUTY, IN
camels_03384450	LUSK CREEK NEAR EDDYVILLE, IL
camels_03439000	FRENCH BROAD RIVER AT ROSMAN, NC
camels_03456500	EAST FORK PIGEON RIVER NEAR CANTON, NC
camels_03460000	CATALOOCHEE CREEK NEAR CATALOOCHEE, NC
camels_03463300	SOUTH TOE RIVER NEAR CELO, NC
camels_03471500	S F HOLSTON RIVER AT RIVERSIDE NR CHILHOWIE, VA
camels_03473000	S F HOLSTON RIVER NEAR DAMASCUS, VA
camels_03479000	WATAUGA RIVER NEAR SUGAR GROVE, NC
camels_03488000	N F HOLSTON RIVER NEAR SALTVILLE, VA
camels_03498500	LITTLE RIVER NEAR MARYVILLE, TN
camels_03500000	LITTLE TENNESSEE RIVER NEAR PRENTISS, NC
camels_03500240	CARTOOGEC HAYE CREEK NEAR FRANKLIN, NC
camels_03504000	NANTAHALA RIVER NEAR RAINBOW SPRINGS, NC
camels_03574500	PAINT ROCK RIVER NEAR WOODVILLE AL
camels_03604000	BUFFALO RIVER NEAR FLAT WOODS, TN
camels_04015330	KNIFE RIVER NEAR TWO HARBORS, MN
camels_04024430	NEMADJI RIVER NEAR SOUTH SUPERIOR, WI
camels_04027000	BAD RIVER NEAR ODANAH, WI
camels_04040500	STURGEON RIVER NEAR SIDNAW, MI
camels_04045500	TAHQUAMENON RIVER NEAR PARADISE, MI
camels_04057510	STURGEON RIVER NEAR NAHMA JUNCTION, MI
camels_04057800	MIDDLE BRANCH ESCANABA RIVER AT HUMBOLDT, MI
camels_04059500	FORD RIVER NEAR HYDE, MI
camels_04063700	POPPLE RIVER NEAR FENCE, WI
camels_04074950	WOLF RIVER AT LANGLADE, WI
camels_04105700	AUGUSTA CREEK NEAR AUGUSTA, MI
camels_04122200	WHITE RIVER NEAR WHITEHALL, MI
camels_04122500	PERE MARQUETTE RIVER AT SCOTTVILLE, MI
camels_04127997	STURGEON RIVER AT WOLVERINE, MI
camels_04185000	Tiffin River at Stryker OH
camels_04196800	Tymochtee Creek at Crawford OH
camels_04197100	Honey Creek at Melmore OH

camels_04213000	Conneaut Creek at Conneaut OH
camels_04216418	TONAWANDA CREEK AT ATTICA NY
camels_04221000	GENESEE RIVER AT WELLSVILLE NY
camels_04224775	CANASERAGA CREEK ABOVE DANSVILLE NY
camels_04256000	INDEPENDENCE RIVER AT DONNATTSBURG NY
camels_04296000	BLACK RIVER AT COVENTRY, VT
camels_05057200	BALDHILL CREEK NR DAZEY, ND
camels_05120500	WINTERING RIVER NR KARLSRUHE, ND
camels_05291000	WHETSTONE RIVER NEAR BIG STONE CITY, SD
camels_05362000	JUMP RIVER AT SHELDON, WI
camels_05393500	SPIRIT RIVER AT SPIRIT FALLS, WI
camels_05399500	BIG EAU PLEINE RIVER AT STRATFORD, WI
camels_05408000	KICKAPOO RIVER AT LA FARGE, WI
camels_05413500	GRANT RIVER AT BURTON, WI
camels_05414000	PLATTE RIVER NEAR ROCKVILLE, WI
camels_05444000	ELKHORN CREEK NEAR PENROSE, IL
camels_05458000	Little Cedar River near Ionia, IA
camels_05466500	EDWARDS RIVER NEAR NEW BOSTON, IL
camels_05487980	White Breast Creek near Dallas, IA
camels_05489000	Cedar Creek near Bussey, IA
camels_05495000	Fox River at Wayland, MO
camels_05495500	BEAR CREEK NEAR MARCELLINE, IL
camels_05501000	North River at Palmyra, MO
camels_05503800	Crooked Creek near Paris, MO
camels_05507600	Lick Creek at Perry, MO
camels_05525500	SUGAR CREEK AT MILFORD, IL
camels_05556500	BIG BUREAU CREEK AT PRINCETON, IL
camels_05584500	LA MOINE RIVER AT COLMAR, IL
camels_05592050	ROBINSON CREEK NEAR SHELBYVILLE, IL
camels_05593575	LITTLE CROOKED CREEK NEAR NEW MINDEN, IL
camels_05593900	EAST FORK SHOAL CREEK NEAR COFFEEN, IL
camels_05595730	RAYSE CREEK NEAR WALTONVILLE, IL
camels_06224000	BULL LAKE CREEK ABOVE BULL LAKE, WY

camels_06280300	SOUTH FORK SHOSHONE RIVER NEAR VALLEY, WY
camels_06289000	Little Bighorn River at State Line nr Wyola MT
camels_06332515	BEAR DEN CREEK NR MANDAREE, ND
camels_06344600	GREEN RIVER NR NEW HRADEC, ND
camels_06350000	CANNONBALL RIVER AT REGENT, ND
camels_06352000	CEDAR CREEK NR HAYNES, ND
camels_06404000	BATTLE CR NEAR KEYSTONE SD
camels_06406000	BATTLE CR AT HERMOSA SD
camels_06409000	CASTLE CR ABOVE DEERFIELD RES NEAR HILL CITY SD
camels_06447500	LITTLE WHITE R NEAR MARTIN SD
camels_06470800	BEAR CREEK NR OAKES, ND
camels_06477500	FIRESTEEL CR NEAR MOUNT VERNON SD
camels_06479215	BIG SIOUX RIVER NR FLORENCE SD
camels_06479438	BIG SIOUX R NEAR WATERTOWN SD
camels_06622700	NORTH BRUSH CREEK NEAR SARATOGA, WY
camels_06623800	ENCAMPMENT RIVER AB HOG PARK CR, NR ENCAMPMENT, WY
camels_06632400	ROCK CREEK AB KING CANYON CANAL, NR ARLINGTON, WY
camels_06803510	LITTLE SALT CREEK NEAR LINCOLN, NEBR.
camels_06803530	ROCK CREEK NEAR CERESCO, NEBR.
camels_06814000	TURKEY C NR SENECA, KS
camels_06853800	WHITE ROCK C NR BURR OAK, KS
camels_06876700	SALT C NR ADA, KS
camels_06878000	CHAPMAN C NR CHAPMAN, KS
camels_06885500	BLACK VERMILLION R NR FRANKFORT, KS
camels_06888500	MILL C NR PAXICO, KS
camels_06889500	SOLDIER C NR TOPEKA, KS
camels_06892000	STRANGER C NR TONGANOXIE, KS
camels_06903400	Chariton River near Chariton, IA
camels_06910800	MARAIS DES CYGNES R NR READING, KS
camels_06917000	L OSAGE R AT FULTON, KS
camels_06918460	Turnback Creek above Greenfield, MO

camels_06921070	Pomme de Terre River near Polk, MO
camels_06921200	Lindley Creek near Polk, MO
camels_07057500	North Fork River near Tecumseh, MO
camels_07060710	North Sylamore Creek near Fifty Six, AR
camels_07066000	Jacks Fork at Eminence, MO
camels_07142300	RATTLESNAKE C NR MACKSVILLE, KS
camels_07145700	SLATE C AT WELLINGTON, KS
camels_07167500	OTTER C AT CLIMAX, KS
camels_07180500	CEDAR C NR CEDAR POINT, KS
camels_07184000	LIGHTNING C NR MCCUNE, KS
camels_07196900	Baron Fork at Dutch Mills, AR
camels_07197000	Baron Fork at Eldon, OK
camels_07208500	RAYADO CREEK NEAR CIMARRON, NM
camels_07261000	Cadron Creek near Guy, AR
camels_07291000	HOMOCHITTO RIVER AT EDDICETON, MS
camels_07299670	Groesbeck Ck at SH 6 nr Quanah, TX
camels_07301410	Sweetwater Ck nr Kelton, TX
camels_07315200	E Fk Little Wichita Rv nr Henrietta, TX
camels_07315700	Mud Creek near Courtney, OK
camels_07335700	Kiamichi River near Big Cedar, OK
camels_07340300	Cossatot River near Vandervoort, AR
camels_07346045	Black Cypress Bayou at Jefferson, TX
camels_07362100	SMACKOVER CREEK NEAR SMACKOVER, ARK.
camels_08013000	CALCASIEU RIVER NR GLENMORA, LA
camels_08014500	Whiskey Chitto Creek Near Oberlin, LA
camels_08023080	Bayou Grand Cane near Stanley, LA
camels_08029500	Big Cow Ck nr Newton, TX
camels_08066200	Long King Ck at Livingston, TX
camels_08066300	Menard Ck nr Rye, TX
camels_08070000	E Fk San Jacinto Rv nr Cleveland, TX
camels_08086212	Hubbard Ck bl Albany, TX
camels_08086290	Big Sandy Ck abv Breckenridge, TX
camels_08101000	Cowhouse Ck at Pidcoke, TX

camels_08104900	S Fk San Gabriel Rv at Georgetown, TX
camels_08109700	Middle Yegua Ck nr Dime Box, TX
camels_08150800	Beaver Ck nr Mason, TX
camels_08158700	Onion Ck nr Driftwood, TX
camels_08164300	Navidad Rv nr Hallettsville, TX
camels_08164600	Garcitas Ck nr Inez, TX
camels_08165300	N Fk Guadalupe Rv nr Hunt, TX
camels_08171300	Blanco Rv nr Kyle, TX
camels_08175000	Sandies Ck nr Westhoff, TX
camels_08189500	Mission Rv at Refugio, TX
camels_08190000	Nueces Rv at Laguna, TX
camels_08190500	W Nueces Rv nr Brackettville, TX
camels_08194200	San Casimiro Ck nr Freer, TX
camels_08196000	Dry Frio Rv nr Reagan Wells, TX
camels_08202700	Seco Ck at Rowe Ranch nr D'Hanis, TX
camels_08269000	RIO PUEBLO DE TAOS NEAR TAOS, NM
camels_08324000	JEMEZ RIVER NEAR JEMEZ, NM
camels_08377900	RIO MORA NEAR TERRERO, NM
camels_08378500	PECOS RIVER NEAR PECOS, NM
camels_08380500	GALLINAS CREEK NEAR MONTEZUMA, NM
camels_09081600	CRYSTAL RIVER AB AVALANCHE C, NEAR REDSTONE, CO.
camels_09210500	FONTENELLE C NR HERSCHLER RANCH, NR FONTENELLE, WY
camels_09223000	HAMS FORK BELOW POLE CREEK, NEAR FRONTIER, WY
camels_09312600	WHITE RIVER BL TABBYUNE C NEAR SOLDIER SUMMIT, UT
camels_09352900	VALLECITO CREEK NEAR BAYFIELD, CO.
camels_09386900	RIO NUTRIA NEAR RAMAH, NM
camels_09404450	EAST FORK VIRGIN RIVER NEAR GLENDALE, UT
camels_09430600	MOGOLLON CREEK NEAR CLIFF, NM
camels_09492400	EAST FORK WHITE RIVER NEAR FORT APACHE, AZ.
camels_09494000	WHITE RIVER NEAR FORT APACHE, AZ.
camels_09497980	CHERRY CREEK NEAR GLOBE, AZ.

camels_09505350	DRY BEAVER CREEK NEAR RIMROCK, AZ.
camels_09505800	WEST CLEAR CREEK NEAR CAMP VERDE, AZ.
camels_09510200	SYCAMORE CREEK NEAR FORT MCDOWELL, AZ.
camels_09512280	CAVE CREEK BLW COTTONWOOD CR NEAR CAVE CREEK, AZ.
camels_09513780	NEW RIVER NEAR ROCK SPRINGS, AZ.
camels_10172700	VERNON CREEK NEAR VERNON, UT
camels_10205030	SALINA CREEK NEAR EMERY, UT
camels_10234500	BEAVER RIVER NEAR BEAVER, UT
camels_10242000	COAL CREEK NEAR CEDAR CITY, UT
camels_10249300	S TWIN R NR ROUND MOUNTAIN, NV
camels_10258500	PALM CYN C NR PALM SPRINGS CA
camels_10259200	DEEP C NR PALM DESERT CA
camels_10263500	BIG ROCK C NR VALYERMO CA
camels_10329500	MARTIN C NR PARADISE VALLEY, NV
camels_10396000	DONNER UND BLITZEN RIVER NR FRENCHGLEN OR
camels_11124500	SANTA CRUZ C NR SANTA YNEZ CA
camels_11143000	BIG SUR R NR BIG SUR CA
camels_11148900	NACIMIENTO R BL SAPAQUE C NR BRYSON CA
camels_11151300	SAN LORENZO C BL BITTERWATER C NR KING CITY CA
camels_11162500	PESCADERO C NR PESCADERO CA
camels_11176400	ARROYO VALLE BL LANG CN NR LIVERMORE CA
camels_11224500	LOS GATOS C AB NUNEZ CYN NR COALINGA CA
camels_11230500	BEAR C NR LAKE THOMAS A EDISON CA
camels_11253310	CANTUA C NR CANTUA CREEK CA
camels_11264500	MERCED R A HAPPY ISLES BRIDGE NR YOSEMITE CA
camels_11266500	MERCED R A POHONO BRIDGE NR YOSEMITE CA
camels_11274500	ORESTIMBA C NR NEWMAN CA
camels_11274630	DEL PUERTO C NR PATTERSON CA
camels_11381500	MILL C NR LOS MOLINOS CA
camels_11383500	DEER C NR VINA CA
camels_11451100	NF CACHE C A HOUGH SPRING NR CLEARLAKE OAKS CA
camels_11468500	NOYO R NR FORT BRAGG CA

camels_11473900	MF EEL R NR DOS RIOS CA
camels_11478500	VAN DUZEN R NR BRIDGEVILLE CA
camels_11480390	MAD R AB RUTH RES NR FOREST GLEN CA
camels_11481200	LITTLE R NR TRINIDAD CA
camels_11482500	REDWOOD C A ORICK CA
camels_11522500	SALMON R A SOMES BAR CA
camels_11523200	TRINITY R AB COFFEE C NR TRINITY CTR CA
camels_11528700	SF TRINITY R BL HYAMPOM CA
camels_11532500	SMITH R NR CRESCENT CITY CA
camels_12010000	NASELLE RIVER NEAR NASELLE, WA
camels_12013500	WILLAPA RIVER NEAR WILLAPA, WA
camels_12020000	CHEHALIS RIVER NEAR DOTY, WA
camels_12025700	SKOOKUMCHUCK RIVER NEAR VAIL, WA
camels_12035000	SATSOP RIVER NEAR SATSOP, WA
camels_12040500	QUEETS RIVER NEAR CLEARWATER, WA
camels_12041200	HOH RIVER AT US HIGHWAY 101 NEAR FORKS, WA
camels_12043000	CALAWAH RIVER NEAR FORKS, WA
camels_12054000	DUCKABUSH RIVER NEAR BRINNON, WA
camels_12056500	NF SKOKOMISH R BL STAIRCASE RPDS NR HOODSPORT, WA
camels_12082500	NISQUALLY RIVER NEAR NATIONAL, WA
camels_12092000	PUYALLUP RIVER NEAR ELECTRON, WA
camels_12115000	CEDAR RIVER NEAR CEDAR FALLS, WA
camels_12143600	SF SNOQUALMIE RIVER AT EDGEWICK, WA
camels_12144000	SF SNOQUALMIE RIVER AT NORTH BEND, WA
camels_12147500	NORTH FORK TOLT RIVER NEAR CARNATION, WA
camels_12167000	NF STILLAGUAMISH RIVER NEAR ARLINGTON, WA
camels_12175500	THUNDER CREEK NEAR NEWHALEM, WA
camels_12186000	SAUK RIVER AB WHITECHUCK RIVER NEAR DARRINGTON, WA
camels_12189500	SAUK RIVER NEAR SAUK, WA
camels_12381400	South Fork Jocko River near Arlee MT
camels_12390700	Prospect Creek at Thompson Falls MT

camels_12411000	NF COEUR D ALENE R AB SHOSHONE CK NR PRICHARD ID
camels_12451000	STEHEKIN RIVER AT STEHEKIN, WA
camels_12488500	AMERICAN RIVER NEAR NILE, WA
camels_13011500	PACIFIC CREEK AT MORAN WY
camels_13011900	BUFFALO FORK AB LAVA CREEK NR MORAN WY
camels_13023000	GREYS RIVER AB RESERVOIR NR ALPINE WY
camels_13083000	TRAPPER CREEK NR OAKLEY ID
camels_13161500	BRUNEAU RIVER AT ROWLAND NV
camels_13235000	SF PAYETTE RIVER AT LOWMAN ID
camels_13240000	LAKE FORK PAYETTE RIVER AB JUMBO CR NR MCCALL ID
camels_13313000	JOHNSON CREEK AT YELLOW PINE ID
camels_13331500	MINAM RIVER NEAR MINAM, OR
camels_14020000	UMATILLA RIVER ABOVE MEACHAM CREEK, NR GIBBON, OR
camels_14096850	BEAVER CREEK, BLW QUARTZ CR, NR SIMNASHO, OR.
camels_14137000	SANDY RIVER NEAR MARMOT, OR
camels_14154500	ROW RIVER ABOVE PITCHER CREEK NEAR, DORENA, OREG
camels_14158500	MCKENZIE RIVER AT OUTLET OF CLEAR LAKE, OR
camels_14166500	LONG TOM RIVER NEAR NOTI, OREG.
camels_14182500	LITTLE NORTH SANTIAM RIVER NEAR MEHAMA, OR
camels_14185000	SOUTH SANTIAM RIVER BELOW CASCADIA, OR
camels_14185900	QUARTZVILLE CREEK NEAR CASCADIA, OREG.
camels_14216500	MUDDY CREEK BELOW CLEAR CREEK NEAR COUGAR, WA
camels_14222500	EAST FORK LEWIS RIVER NEAR HEISSON, WA
camels_14236200	TILTON RIVER AB BEAR CANYON CREEK NEAR CINEBAR, WA
camels_14301000	NEHALEM RIVER NEAR FOSS, OR
camels_14305500	SILETZ RIVER AT SILETZ, OR
camels_14306500	ALSEA RIVER NEAR TIDEWATER, OR
camels_14309500	WEST FORK COW CREEK NEAR GLENDALE, OR
camels_14316700	STEAMBOAT CREEK NEAR GLIDE, OR
camels_14325000	SOUTH FORK COQUILLE RIVER AT POWERS, OR
camels_14400000	CHETCO RIVER NEAR BROOKINGS, OR

The distribution of the gauge stations associated with the selected catchments across CONUS is depicted in Fig C 1. In Table C 2, the number of selected catchments associated with each HU is shown.



Fig C 1. The catchments' gauge station distribution across the CONUS.

Table C 2. Number of selected catchments in each HU.

HU	Count
1	16
2	50
3	57
4	22
5	27
6	14
7	24
8	4
9	2
10	34
11	19
12	23
13	5
14	5
15	11
16	6
17	54
18	29

C2 Formulation

Long Short-Term Memory Network

The formulation of a LSTM cell associated with timestep T is as follows (Jahangir et al., 2023; Kratzert et al., 2019):

$$\mathbf{i}^T = \text{sigmoid}(\mathbf{W}_i \mathbf{X}^T + \mathbf{U}_i \mathbf{h}^{T-1} + \mathbf{b}_i) \quad \text{Eq C 1}$$

$$\mathbf{f}^T = \text{sigmoid}(\mathbf{W}_f \mathbf{X}^T + \mathbf{U}_f \mathbf{h}^{T-1} + \mathbf{b}_f) \quad \text{Eq C 2}$$

$$\mathbf{g}^T = \tanh(\mathbf{W}_g \mathbf{X}^T + \mathbf{U}_g \mathbf{h}^{T-1} + \mathbf{b}_g) \quad \text{Eq C 3}$$

$$\mathbf{o}^T = \text{sigmoid}(\mathbf{W}_o \mathbf{X}^T + \mathbf{U}_o \mathbf{h}^{T-1} + \mathbf{b}_o) \quad \text{Eq C 4}$$

$$\mathbf{c}^T = \mathbf{f}^T \odot \mathbf{c}^{T-1} + \mathbf{i}^T \odot \mathbf{g}^T \quad \text{Eq C 5}$$

$$\mathbf{h}^T = \mathbf{o}^T \odot \tanh(\mathbf{c}^T) \quad \text{Eq C 6}$$

where $\text{sigmoid}(x) = 1/e^{-x}$, $\tanh(x) = (e^{2x} - 1)/(e^{2x} + 1)$, and $\mathbf{g}^T$, $\mathbf{h}^T$, $\mathbf{c}^T$ are the cell input, recurrent state, and cell state at timestep T , respectively. The $\odot$ operator indicates the element-wise multiplication. The final recurrent state is passed to the dense network.

Error Metrics

The formulae for the error metrics used to evaluate the different models include the normalized root mean square (nRMSE), normalized mean absolute error (nMAE; see Quilty et al. (2023)), modified Kling–Gupta efficiency (Kling et al., 2012), and Nash-Sutcliffe efficiency (NSE; McCuen et al., 2006):

$$\text{nRMSE} = \sqrt{1/N \sum_{i=1}^N (\hat{Q}_i - Q_i)^2} / \mu_Q \quad \text{Eq C 7}$$

$$\text{KGE} = 1 - \sqrt{(\rho - 1)^2 + \left(\frac{\sigma_{\hat{Q}}}{\sigma_Q} - 1\right)^2 + \left(\frac{\mu_{\hat{Q}}}{\mu_Q} - 1\right)^2} \quad \text{Eq C 8}$$

$$\text{NSE} = 1 - \frac{\sum_{i=1}^N (\hat{Q}_i - Q_i)^2}{\sum_{i=1}^N (\mu_Q - Q_i)^2} \quad \text{Eq C 9}$$

$$\text{nMAE} = 1/N \sum_{i=1}^N |\hat{Q}_i - Q_i| / \mu_Q$$

Eq C 10

where μ and σ are the mean and standard deviation operators, respectively. The parameter N indicates the total number of samples and ρ is the linear correlation (Pearson correlation) between forecasts and observations.

C3 Daily Forecast Performance Across CONUS

The distribution of the obtained error metrics across CONUS for the normalizing flow (NF) and the deterministic counterpart model (DT) is presented in Fig C 2. The mean is calculated across all seven timesteps and all selected catchments.

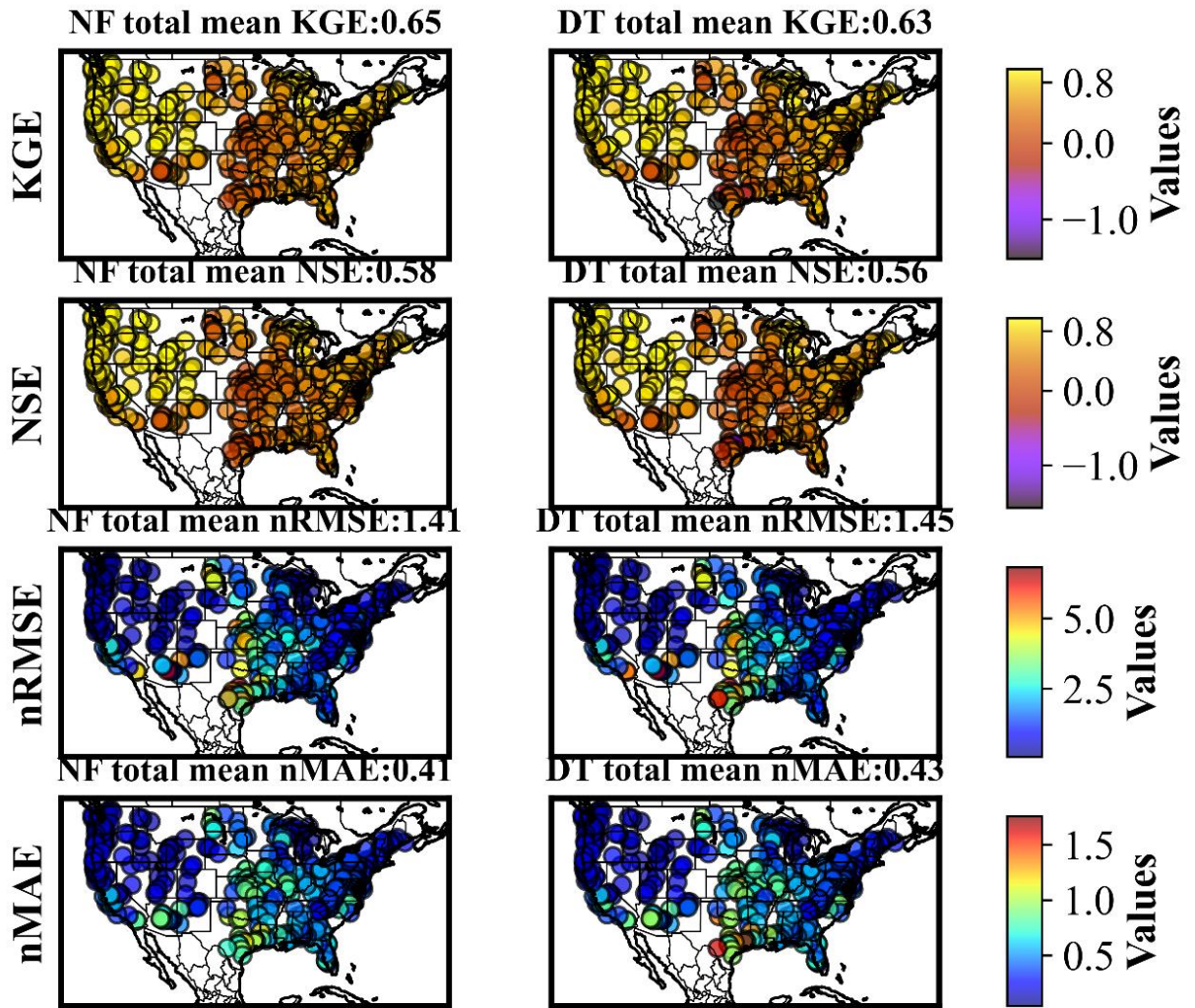


Fig C 2. Distribution of error metrics across CONUS.

C4 Comparison Between ERA5 Land and Daymet

Performance comparison between two distinct LSTM models utilizing Daymet and ERA5 Land data for same-day streamflow prediction (simulation). The LSTM model was comprised of a LSTM layer, with two dense layers on top, where the final layer was used to estimate the parameters of a mixture Gaussian (MG) network (see Klotz et al. (2022) for details). An input lag (look-back period) of 365 days was adopted. The LSTM and the first dense layer size, dense layer activation type, and number of Gaussian components were selected using Bayesian optimization. Please refer to the main text Section 4.4.3 for details.

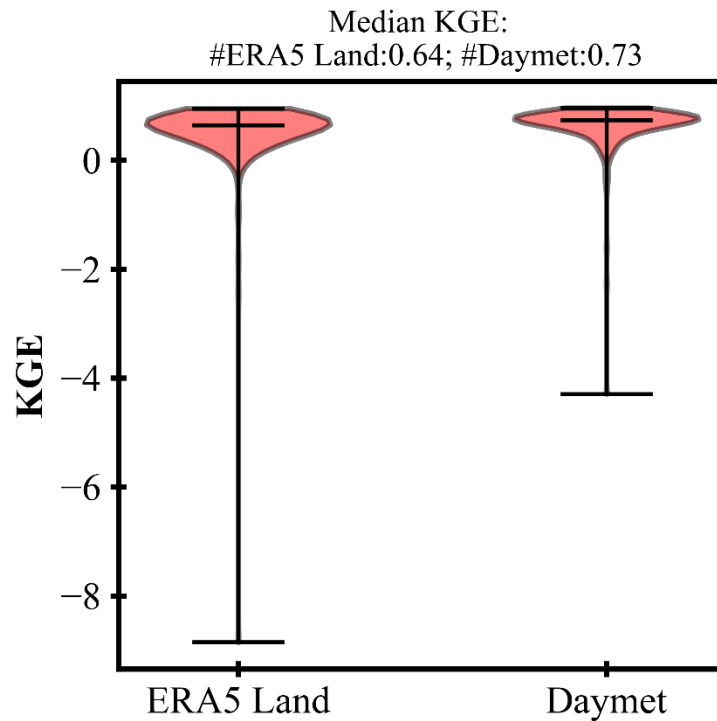


Fig C 3. Simulation model performance comparison using different input types.

C5 Weekly Performance Assessment (Non-Reconciled Forecasts)

In Table C 3, the weekly forecast performance is provided.

Table C 3. Weekly performance of the NF model. Metrics were calculated across all selected catchments.

Metric	Median	Mean
KGE	0.75	0.65
NSE	0.69	0.63
nRMSE	0.69	1.32
nMAE	0.31	0.44