

Upscaling and downscaling snow processes with machine learning in watershed models

by

Hannah Burdett

A thesis

presented to the University of Waterloo

in fulfilment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Civil Engineering (Water)

Waterloo, Ontario, Canada, 2025

© Hannah Burdett 2025

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:

Bart Nijssen
Professor
Department of Civil and Environmental Engineering
University of Washington

Supervisor:

James R. Craig
Professor
Department of Civil and Environmental Engineering
University of Waterloo

Internal Members:

Bryan Tolson
Professor
Department of Civil and Environmental Engineering
University of Waterloo

Juliane Mai
Research Associate Professor
Earth and Environmental Sciences
University of Waterloo

Internal-External Member:

Christopher Fletcher
Associate Professor
Geography and Environmental Management
University of Waterloo

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

Hannah Burdett was the sole author of Chapters 1 and 6, written under the supervision of Dr. James R. Craig. Chapters 3, 4, and 5 were prepared for publication, with Chapter 3 currently under review in *Water Resources Research* (Burdett and Craig, in review). Exceptions to sole authorship of material are as follows:

Chapter 3 and 5 research was conducted by Hannah Burdett under the supervision of Dr. James R. Craig. Dr. Craig contributed to the design of the study and provided iterative feedback during the methodology development and analysis phase. Hannah Burdett carried out all other aspects of the study and drafted the manuscript, which Dr. Craig reviewed and edited.

Chapter 4 research was conducted by Hannah Burdett under the supervision of Dr. James R. Craig, in collaboration with Matt Chernos, Dr. Ryan MacDonald, and Dr. Chris Hopkinson. Matt Chernos and Dr. MacDonald developed and calibrated the hydrologic model for the Oldman River Basin, which was later updated by Hannah Burdett. Dr. Hopkinson provided support through the Alberta Innovates Project, while Dr. Craig offered key insights during manuscript preparation. Hannah Burdett carried out all other aspects of the study.

Abstract

Hydrologic models play a vital role in understanding and predicting the movement of water within watersheds, providing essential insights for effective management and sustainability of water resources. However, watersheds exhibit significant heterogeneity in their landscape properties and complex responses to spatiotemporal variations in climatic inputs. This variability introduces a gap between the representation of physical processes at the point scale and their behaviour at the watershed scale, making it challenging to accurately capture the full complexity of the hydrologic cycle across different spatial scales. Bridging this gap requires the identification of effective scaling approaches tailored to capture the complexities across scales. Scaling approaches look to translate information from one scale to another, whether moving from a smaller to a larger scale (upscaling) or from a larger to a smaller scale (downscaling). Although various approaches in the literature have been applied to develop scaling methods for forcing variables, such as precipitation and temperature, and fluxes (e.g., evapotranspiration), there is a notable gap in deriving and applying scaling techniques for snow-related variables, such as SWE, snowmelt, or sublimation. Addressing this gap may help in improving hydrologic model accuracy in snow-dominated regions, where snow dynamics significantly influence water availability and watershed resources.

The primary objective of this thesis is to develop, implement, and evaluate machine learning-based upscaling methodologies to aid in understanding the relationship between local-scale snow-related variables, landscape heterogeneity, and the large-scale hydrologic response of a catchment. Such methods are useful for effectively simulating the net impact of local variability in snow processes without resorting to fine-resolution models. A secondary focus of this research aims to identify the conditions under which emergent constitutive relationships specific to snow-related fluxes are (or are not) valid and to assess the transferability of these relationships. Finally, this work introduces a machine learning-based downscaling approach that refines large-scale mean model outputs into localized snow states and fluxes. Together, these scaling techniques explore the potential of machine learning to address challenges in hydrologic scaling specific to snow-related fluxes.

Acknowledgements

I would like to sincerely thank Dr. James R. Craig for his guidance, insight, and support throughout this Ph.D. journey. His encouragement, artistic talent, and sense of humor have made this process both intellectually enriching and genuinely enjoyable. It has been a privilege to learn from and work alongside such a brilliant supervisor.

I want to acknowledge my PhD committee, including Bart, Julie, Chris, and Bryan, for taking the time to review my thesis and providing their invaluable feedback.

I would like to express my appreciation to my friends and colleagues at the University of Waterloo for their support, insightful discussions, and encouragement over the past years. Particularly Larissa for the countless chats that kept me grounded, the coffee breaks that kept me going, and for always opening your home (and couch) whenever I needed a place to stay.

Furthermore, I would like to thank Haydn, Holden, Dean, and Alicia. I would not be the person I am without all of you. Whether it was encouraging me to take a break, offering a much-needed laugh in moments of stress, or simply being there no matter what. I am endlessly grateful for you all.

Last but certainly not least, I want to express my deepest gratitude to Jack. Through all the time apart, the late-night work sessions, and the moments of doubt, you have stood by me with unwavering patience, love, and encouragement. From the very beginning, you have been my greatest source of support, and in many ways, you deserve honorary co-authorship on this work. I couldn't ask for a better partner.

Dedication

This thesis is dedicated to my four-legged study buddy, Lilly, who has been by my side through high school, undergrad, my master's, and now my Ph.D. Your unwavering enthusiasm for walk breaks, impeccable timing in interrupting Zoom meetings, and expert-level cuddling skills have made this journey infinitely better.

Table of Contents

List of Tables	x
List of Figures	xi
List of Abbreviations.....	xiii
1 Introduction	1
1.1 Thesis Objectives.....	2
2 Literature Review.....	4
2.1 Hydrological Models.....	4
2.2 Heterogeneity and Spatial Scale	5
2.3 Upscaling in Hydrologic Modelling	7
2.4 Constitutive Relationships	8
2.5 Downscaling in Hydrologic Modelling	9
2.6 Machine Learning in Hydrology	11
2.7 Scaling with Machine Learning.....	14
2.8 Snow Processes	16
3 Upscaling snow accumulation and ablation using machine learning.....	20
3.1 Introduction.....	20
3.2 Machine-learning Upscaling Methodology Framework.....	23
3.2.1 The Fine Resolution Model	24
3.2.2 Training the Machine Learning Model	28
3.3 Results and Discussion	33
3.3.1 Validation	33
3.3.2 Single HRU (Homogeneous)	35
3.3.3 Random Stratified Spatial Validation (Heterogeneous Dissimilar)	37
3.3.4 Similar HRU (Heterogeneous Similar)	41
3.4 Conclusion	46
4 Assessing machine learning-identified upscaled constitutive relationships for snow accumulation and ablation in mountainous terrain.....	48
4.1 Introduction.....	48
4.2 Methods	51
4.2.1 Study Area	51

4.2.2	Datasets.....	52
4.2.3	Fine-resolution Hydrologic Model	53
4.2.4	General Experimental Design	53
4.2.5	Overview of Experiments	56
4.2.6	Machine Learning Algorithm Comparison.....	56
4.2.7	Spatial Resolution Comparison	59
4.2.8	Variable Reduction	59
4.2.9	Transferability of ML Relationships.....	61
4.3	Results and Discussion	62
4.3.1	Machine Learning Algorithm Comparison.....	62
4.3.2	Spatial Resolution Comparison	65
4.3.3	Variable Reduction	67
4.3.4	Relationship Transfer.....	71
4.4	Conclusion	76
5	Downscaling snowpack evolution processes with machine learning.....	78
5.1	Introduction.....	78
5.2	Methods	80
5.2.1	Study Area and Data	80
5.2.2	Machine Learning Downscaling Methodology	81
5.2.3	Fine-Resolution Model.....	82
5.2.4	Scaling Ratios.....	83
5.2.5	Machine Learning Algorithm.....	83
5.3	Results and Discussion	87
5.3.1	Spatial Validation.....	87
5.3.2	Temporal Fine Scale Reconstruction.....	96
5.3.3	Validation of Upscaling and Downscaling Methodologies.....	101
5.4	Conclusion	103
6	Conclusions	105
6.1	Summary.....	105
6.2	Contributions.....	106
6.3	Limitations and Future Work.....	107
	References.....	111

List of Tables

Table 3.1: KGE values for each of the subbasins when comparing the Raven estimated daily SWE values against the Daymet SWE observations.....	34
Table 4.2: An overview of the ML algorithm comparison setup, outlining the architectures of each ML model, clustering configuration, domain specifications, included distributional information, and the error metrics used for evaluation.	58
Table 4.3: Outlines the setup for the spatial resolution comparison experiment using an RF model.	59
Table 4.4: Overview of the Variable Reduction experiment outlining the ML model used, the number of clusters and domains, the inclusion of distributional information, and evaluation criteria that are used. 60	
Table 4.5: Overview of the Transferability of ML Relationships experiment outlining the ML model used, the number of clusters and domains, the inclusion of distributional information, and evaluation criteria that are used.....	62

List of Figures

Figure 2.1: Conceptual illustration of upscaling and downscaling processes, inspired by Blöschl and Sivapalan (1995).	7
Figure 2.2: Schematic of a basic feedforward multilayer perceptron architecture, where the input values (grey circles) pass through multiple hidden layers, where each neuron processes the inputs with weights, biases, and activation functions. The final result (pink circle) is produced in the output layer after processing, with the network adjusting weights and biases to minimize error and improve predictions. ..	12
Figure 2.3: The random forest model selects features and builds a forest of decision trees by splitting the initial values into subsets (represented by grey circles) for each tree. The mean squared error is then calculated independently for each subset, with the blue circles representing the values that result in the lowest mean squared error.....	13
Figure 3.1: Conceptual outline of the steps used in the machine-learning based upscaling methodology.	24
Figure 3.2: The final discretization of the Kananaskis River Basin, defined upstream of the 05BF025 hydrometric flow gauge (51°02'30" N, 115°01'55" W), resulted in 20,000 HRUs. The larger blue areas represent lakes, while the numbers correspond to the subbasin identification numbers used in BasinMaker.	26
Figure 3.3: The contents of the input and output vector for the random forest model. All mean forcings are spatial averages over the specified computational scale. Input and output vector terms with an asterisk indicate that the input is optionally used in some subset of experiments.	29
Figure 3.4: A plot of the 12 sections of the Kananaskis River Basin divided along the longitude and corresponding table indicating the number of HRUs in each section and elevation in each section.....	32
Figure 3.5: Histograms compare the distribution of the SWE values for the fine-resolution Raven and Daymet datasets. Additionally, probability density function curves of the non-zero values are overlaid to further assist in visualizing fit of the datasets.....	35
Figure 3.6: Modified KGE score CDFs of the 4,000 HRUs from the Homogeneous experiment for each of the three model instances.	36
Figure 3.7: Box plot of the Heterogeneous Dissimilar KGE values for each section evaluating the level of agreement between the high-resolution Raven emulated and RF estimated sublimation, snowmelt, SCA, and mass balance estimated SWE.	38
Figure 3.8: Mean SCA, snowmelt, sublimation, and SWE in section 6 comparing the outputs of the Raven model and mass balance with and without inclusion of distributional information over 3000 daily timesteps.	39
Figure 3.9: Box plot of the Heterogeneous Dissimilar KGE values for each section evaluating the level of agreement between the high-resolution Raven emulated and RF estimated sublimation, snowmelt, SCA, and mass balance estimated SWE. Comparisons are shown for RF models using either 6 or 100 trees.....	40
Figure 3.10: A plot of the 12 k-means clusters of the Kananaskis River Basin and a corresponding table indicating the number of HRUs in each domain, elevation, and landcover percentages in each cluster.	42
Figure 3.11: Illustrates the KGE values for all clusters, evaluating the agreement across various variables, including high-resolution Raven emulated SWE, Raven naïvely upscaled SWE, and RF-estimated snowmelt, sublimation, SCA, and mass balance estimated SWE in both Heterogeneous Similar and Dissimilar scenarios with and without distributional information.	43
Figure 3.12: A plot of 3000 daily timesteps of cluster 8 outputs for mean SWE comparing the outputs of the Raven fine-resolution model, naïve upscaled model, and mass balance without and with distributional information.	45
Figure 4.1: The geographic location of the Kananaskis and Oldman watersheds in Alberta, Canada, highlighting the topographic relief of the study area. The bar chart inset illustrates the percentage distribution of the landcover types within each watershed.	52
Figure 4.2: Provides a conceptual representation of the spatial subdivisions used in this study, showing the k-means cluster divided into M domains, each composed of multiple HRUs.....	54

Figure 4.3: A plot of the 12 k-means clusters of the Kananaskis River Basin and a corresponding table indicating the number of HRUs in each section, elevation, and landcover percentages in each cluster.	63
Figure 4.4: Illustrates the KGE, MBE, and R^2 values for all clusters, evaluating the agreement across various variables, including Raven naïvely upscaled, RF, XGBoost, and ANN-estimated snowmelt, sublimation, SCA, and mass balance estimated SWE agreement with the aggregated high-resolution Raven estimates.	64
Figure 4.5: KGE values for mass-balance estimated SWE, snowmelt, and sublimation compared to fine-resolution Raven model snow-related fluxes as the number of domains within each of the 12 clusters was systematically adjusted (4, 8, 16, 32, and 64).	66
Figure 4.6: A plot displaying the significant features identified by the LOO method and KGE values for both the full-variable and LOO method RF models across each cluster. Acronyms: SWR-SD (Shortwave Radiation Standard Deviation), ST (Surface Temperature), SWR (Shortwave Radiation), SCA (Snow-Covered Area), SF (Snowfall), SWE (Snow Water Equivalent), SWE-SD (Snow Water Equivalent Standard Deviation), PRCP (Precipitation), Wind (Wind Speed), Temp (Temperature).	68
Figure 4.7: A plot of the k-means clusters of the Oldman River Basin and a corresponding table indicating the number of HRUs in each section, elevation, and landcover percentages in each cluster.	72
Figure 4.8: KGE, MBE, and R^2 values for all clusters, evaluating the agreement across various variables, including Raven’s naïvely upscaled estimates and the RF-transferred relationships from Kananaskis to the Oldman River Basin.	73
Figure 4.9: Presents a time series comparison of Raven fine-resolution SWE, Oldman mass balance estimated SWE from the RF model, and the naïve approach.	74
Figure 4.10: Presents a time series comparison of Raven fine-resolution SWE, Oldman mass balance estimated SWE from the RF model, and the naïve approach.	75
Figure 5.1: Conceptual outline of the steps used in the machine-learning based downscaling methodology.	82
Figure 5.2: The input and output vectors for the XGBoost model. All mean forcing represents spatial averages over the coarse scale.	85
Figure 5.3: Visualization of the 12 k-means clusters in the Kananaskis River Basin, which are the same clusters used in Chapter 4 accompanied by a table detailing the number of HRUs per cluster, elevation distribution, and landcover percentages.	86
Figure 5.4: Illustrates the KGE, R^2 , and MAE values for the 12 clusters, assessing the agreement between the ML-downscaled SWE values and the original Raven fine-resolution estimates in the Spatial Validation (Ideal) experiment.	89
Figure 5.5: Time series plot comparing fine-resolution Raven SWE, coarse-resolution SWE (before downscaling), and ML-downscaled SWE for three HRUs in clusters 3, 5, and 12. The plot illustrates overall agreement, highlighting differences between a low, moderate, and high performing HRU and a high-performing HRU.	90

List of Abbreviations

ANN – Artificial Neural Network
CDF – Cumulative Distribution Functions
DT – Decision Trees
ET – Evapotranspiration
GCM – Global Climate Model
HRU – Hydrological Response Unit
KGE – Kling-Gupta Efficiency
LOO – Leave-One-Out
LSTM – Long Short-Term Memory
MAE – Mean Absolute Error
MBE – Mean Bias Error
ML – Machine Learning
MSE – Mean Squared Error
PDF – Probability Distribution Function
PRCP – Precipitation
R² – Coefficient of Determination
RCM – Regional Climate Model
RF – Random Forest
SCA – Snow-Covered Area
SCD – Snow Depletion Curve
SF – Snowfall
SHAP – SHapley Additive exPlanations
ST – Surface Temperature
SWE – Snow Water Equivalent
SWE-SD – Snow Water Equivalent Standard Deviation
SVR – Support Vector Regression
SWAT – Soil and Water Assessment Tool
SWR – Shortwave Radiation
SWR-SD – Shortwave Radiation Standard Deviation
XGBoost – Extreme Gradient Boosting

Chapter 1

Introduction

Hydrologic processes exhibit variability across spatial scales, complicating their scientific analysis through both measurements and modelling. In the field, observations of hydrologic processes are typically limited to point measurements or small areas spanning only a few square meters. From these in-field observations, mathematical relationships are developed to describe the physical phenomena of snow accumulation and ablation, which are then integrated into hydrologic models. These models, grounded in small-scale physics, then extend these mathematical relationships to regional or watershed scales, often assuming that the small-scale behaviors are applicable beyond their original point of observation. However, at the scales of our computational units (hydrologic response units, catchments, or watersheds) spatial heterogeneity in landcover, vegetation, soil properties, and topography is ubiquitous. The heterogeneity of the natural environment means that different hydrologic processes occur across a hierarchy of spatial and temporal scales. To simulate this variability across scales, requires understanding the system heterogeneity at the small-scale and deriving defensible simplifications of this heterogeneity that can be applied to larger scales, which is the fundamental question of upscaling. Upscaling has been successfully applied throughout the hydrologic literature to generate large-scale descriptions of small-scale processes (e.g., Klemeš, 1983; Blöschl and Sivapalan, 1995). However, the application of these scaling methods to snow-related fluxes remains less explored. Snow processes, such as melt and sublimation, exhibit distinct spatial patterns that are highly sensitive to sub-computational scale topographic and meteorological variations (Crumley et al., 2024), necessitating targeted scaling approaches to accurately represent snow dynamics across scales within hydrologic models. Here, we address the problem using a bottom-up scaling approach: training machine learning (ML) models to represent coarse-scale snow ablation and accumulation in heterogeneous domains using the results from detailed fine-resolution hydrological simulations.

Recent applications of ML models in catchment hydrology have increased in popularity, driven by advances in computational power and the availability of hydrologic data, from remote sensing imagery to the expansion of in situ sensor technologies (Yuan et al., 2020). Traditional knowledge of hydrologic processes is often insufficient to capture the full range of observed variability in these data, which arises from the nonlinear nature of hydrologic systems (Hispey et al., 2015). The hydrologic literature demonstrates the effectiveness of the ML algorithms in diverse applications, such as assimilating alternative data sources to enhance forecast accuracy (Feng et al., 2020), estimating snow density from remote sensing products (Broxton et al., 2019), superior hydrological forecasting (Kratzert et al., 2021), and generating high-resolution landcover classifications (Geng et al., 2015). The success of ML in hydrology is largely due to its ability to learn directly from data, allowing it to capture complex, nonlinear relationships. ML algorithms, particularly physics-informed and interpretable ML, offer the potential to incorporate physical constraints and provide insights into the underlying processes. ML models hold potential for enhancing spatial scaling in hydrology due to their ability to learn scaling relationships directly from data, without relying on rigid assumptions about the underlying processes. This flexibility may allow ML to excel at both downscaling and upscaling tasks, effectively capturing spatial variability across different scales and potentially generating hydrological relationships which are more transferable than those predicated on the use of an equivalent homogeneous domain. The full potential of ML for upscaling in hydrology remains underexplored, representing a gap in current research. This thesis aims to address this gap by investigating how ML can be leveraged to derive upscaling constitutive relationships and apply a similar ML framework to downscaling.

1.1 Thesis Objectives

This thesis investigates the use of ML methods informed by the output of high-resolution hydrological models to represent emergent constitutive relations between coarse-resolution system states and fluxes in heterogeneous watersheds which may be applied at a relatively large scale (e.g., 5-20km²) without the need for detailed simulation. We first apply an ML upscaling methodology to emulate the spatially-aggregated outputs generated by a fine-scale model. We then query the ML model to gain insights into the scaling relationships between fluxes, system states, and properties. By applying this approach across multiple scales, this study aims to assess the viability of ML algorithms for upscaling and downscaling snow-related processes. In addition, this thesis also applies to downscaling, where ML methods are used to replicate fine-scale variability of hydrologic fluxes and states using the outputs from coarse-scale models.

The objectives of this thesis are to:

1. Develop, apply, and test a ML-based upscaling methodology capable of replicating the spatially averaged states and fluxes (i.e., sublimation, snowmelt, etc.) generated by a high-resolution snow model in mountainous terrain at multiple spatial scales.

2. Mine the results of the ML upscaling model to assess the relative impact of different scale-dependent terrain, snowpack, and forcing characteristics on the hydrologic functioning of a watershed, i.e., using the ML model to discover i) the conditions under which emergent constitutive relations are (or are not) valid, and ii) what the key controls are on upscaling.
3. Extend and apply the ML-based methodology to downscaling of snow related state and fluxes and apply the ML downscaling relationships to results generated from the upscaling methodology.

The net contribution of these methods are ML-informed hydrologic models which can be run at coarse computational resolutions way larger than the scale of important heterogeneity, but may also yield high-resolution reconstructions of states (snow water equivalent and snow covered area) as well as ablation fluxes (sublimation and snowmelt) via downscaling. It is hoped that similar techniques may later be applicable to other components of the hydrologic cycle. For example, these methodologies could be extended to upscale or downscale other hydrologic processes, such as evapotranspiration or infiltration, particularly in regions where fine-scale observations are limited.

To present the methods and findings of this research, the thesis is structured as follows: Chapter 2 includes a comprehensive literature review, covering hydrological models, upscaling and downscaling methods for hydrological processes with a focus on snow, and the application of ML in hydrology and scaling methodologies. Chapter 3 introduces the development of the ML upscaling methodology and the application of this algorithm in a drainage basin in the Canadian Rockies. Chapter 4 investigates the transferability of machine learning-derived constitutive relationships between study areas with similar characteristics, aiming to identify key controls on snow-related variables across scales and assess how these controls impact the scaling capabilities of ML algorithms. Chapter 5 introduces an ML-based downscaling methodology that generates scaling ratios to refine large-scale model outputs into localized snow-related variables. Chapter 6 concludes with a summary of the contributions made.

Chapter 2

Literature Review

2.1 Hydrological Models

Hydrologic models are mathematical equations derived from observations of natural phenomena, designed to simulate the physical processes of hydrology (Singh and Woolhiser, 2002). These models apply physical laws to describe the redistribution of mass and energy within watersheds, supporting the simulation of streamflow, soil moisture, snowpack, and other components of the water cycle. Hydrologic models vary in complexity, ranging from simple empirical models that rely on historical data, to advanced physically-based models that use detailed equations and can account for spatial variability to capture the complex interactions within the hydrologic cycle. These models are essential for understanding and predicting the water cycle, aiding in water resource management, flood forecasting, drought mitigation, and assessing the impacts of climate change and human activities on water availability and quality (Kumar et al., 2023).

Hydrologic models have evolved significantly over time, driven by advancements in computational power and data availability (Archfield et al., 2015). These developments have enabled the creation of increasingly complex and comprehensive models capable of simulating a wide range of hydrological processes under diverse conditions. In recent decades, hydrological models have diversified in both design and application, falling broadly into two categories: conceptual models and physically-based models. Conceptual models represent the aggregate effects of dominant processes through an integrated response, such as catchment-scale streamflow. Rather than relying on detailed physical equations, these models group multiple physical processes into relatively simple mathematical functions (e.g., Burnash et al., 1973; Lindström et al., 1997; Perrin et al., 2003). In contrast, physically-based models explicitly represent dominant processes deterministically, employing detailed mathematical equations grounded in fundamental physical laws to describe and predict hydrological processes

(Refsgaard, 1997). These models typically describe the hydrological cycle using a series of 0- or 1-dimensional water balance equations, each representing the influence of specific processes (e.g., runoff, infiltration) on individual storage components (e.g., topsoil).

Physically-based hydrologic models are further categorized into three general types based on the degree of spatial detail retained: lumped, semi-distributed, and distributed (Golden et al., 2017). Lumped models treat the entire watershed as a single homogeneous unit, with no spatial discretization of input data or model parameters (Clarke, 1973). In contrast, discretized models divide the watershed spatially, with each cell representing a specific location (Kampf & Burges, 2007). This spatial discretization allows for a more detailed understanding of internal processes by calculating changes in mass and energy across a grid of cells. The Freeze-Harlan (1969) blueprint underpins these models, envisioning the world as a finely discretized grid where constitutive laws, parameters, and states are known and appropriately scaled. However, the success of these models depends on accurately parameterizing both state variables and model parameters, as well as ensuring that constitutive laws hold at these scales. This blueprint does not directly address the potential disconnect between the scale of simulation and the scale of physical processes. Semi-distributed models, which blend elements of lumped and distributed models, are spatially discretized but are less explicit about location than fully distributed models. In these models, heterogeneity is often addressed through either implicit or explicit upscaling methods.

Physically-based, spatially distributed models are designed to provide detailed predictions of state variables and fluxes across a watershed, potentially offering a high level of realism and theoretical rigor (McDonnell et al., 2007). Despite their potential, these models face several challenges. A primary issue is that their theoretical foundation often originates from small-scale physics or theories, which do not always scale up effectively to larger systems (Kirchner, 2006). This limitation can lead to inaccuracies when applying these models to larger or more complex watersheds, where the spatial variability of landscape properties and the nonlinear nature of many hydrological processes become more pronounced. Additionally, the inherent complexity and variability of natural landscapes introduce significant difficulties in parameterizing these models accurately. As a result, different parameter combinations can produce similar model outputs, a phenomenon known as equifinality, which Beven (2000) identifies as a major source of prediction uncertainty. This uncertainty is further compounded by the challenge of representing the true spatial heterogeneity of a watershed, potentially limiting the applicability and reliability of physically-based distributed models at larger scales.

2.2 Heterogeneity and Spatial Scale

The heterogeneity of hydrologic processes occurs across many orders of magnitude in spatial scale, from pore-scale flow processes to global cycles of water and energy (Skøien and Blöschl, 2003). The need for scaling arises from this heterogeneity, as processes operate at different scales, and the range of

scales for any given process can be vast (Dehotin and Braud, 2008; Klemeš, 1983). For example, the seasonal development of snowpack is a time-integrative, spatially varying process with distinct periods of accumulation and melt. However, these processes do not occur uniformly across landscapes. Factors like topography and landcover influence snow distribution, with snow accumulation often deeper on north-facing slopes or in forested areas due to shading, while melt rates tend to be higher in open areas or on south-facing slopes due to increased solar exposure.

Measurement techniques, including in-field point-scale observations and remote sensing, are crucial for understanding hydrologic phenomena at different scales (Blöschl and Sivapalan, 1995; Klemeš, 1983). In-field samples, which provide highly detailed, site-specific data, are often limited and typically concentrated in populated areas, making comprehensive monitoring across entire watershed systems impractical (Albert, 2011). This limitation is especially pronounced at high latitudes, where meteorological stations are sparse and generally located at lower elevations, despite critical hydrologic processes, like snow accumulation, being more prominent at higher altitudes (Wang et al., 2019). While ground-based observations and modelling studies provide valuable and localized insights into hydrologic processes, scaling these findings to larger regions remains a significant challenge.

Remote sensing techniques have revolutionized hydrologic observation by enabling direct measurement of various components of the water balance across a wide range of spatial and temporal resolutions (Gao et al., 2018). The use of remote sensing in hydrology can generally be divided into three categories: simple delineation of broad surface features, such as snow cover; classification of features like landcover types; and estimation of hydrological state variables, such as soil moisture, by correlating measured conditions with remote sensing data. However, the spatial resolution of remote sensing data does not always align with the heterogeneous nature of hydrologic processes. Typically, the spatial resolution of hydrologic models is finer than most remote sensing data, often requiring sub-kilometer to meter-scale resolutions to accurately capture localized hydrologic processes. As a result, remote sensing data often require additional processing to match the finer resolution of these models.

Scale issues are not only relevant in monitoring but also in computational hydrology, where the model scale plays a critical role in determining the resolution at which hydrologic processes are simulated. The model scale refers to the area of the smallest computational unit, such as a grid cell or hydrological response unit (HRU). In a discretized model, an HRU is a spatial unit where the hydrologic response to external forcing is considered uniform (even if sub-HRU heterogeneity is known to be present). Thus, scaling is essential in computational hydrology, as it bridges the gap between the scale at which hydrologic phenomena are observed and the scale at which they are represented in models.

In the literature, scaling methods are generally categorized into two types: downscaling and upscaling (Blöschl and Sivapalan, 1995). Downscaling involves reconstructing mesoscale or microscale processes from macroscale data. This typically requires using statistical or dynamic models to translate coarse-resolution information into finer-resolution outputs, effectively capturing local variability and

complex hydrological processes at smaller scales. Upscaling, conversely, refers to transferring detailed microscale information to the macroscale (Anderson et al., 2003). Both upscaling and downscaling are essential for bridging the gap between different spatial scales, allowing for a more comprehensive understanding and modelling of hydrological processes. Figure 2.1 provides a conceptual illustration of the downscaling and upscaling processes.

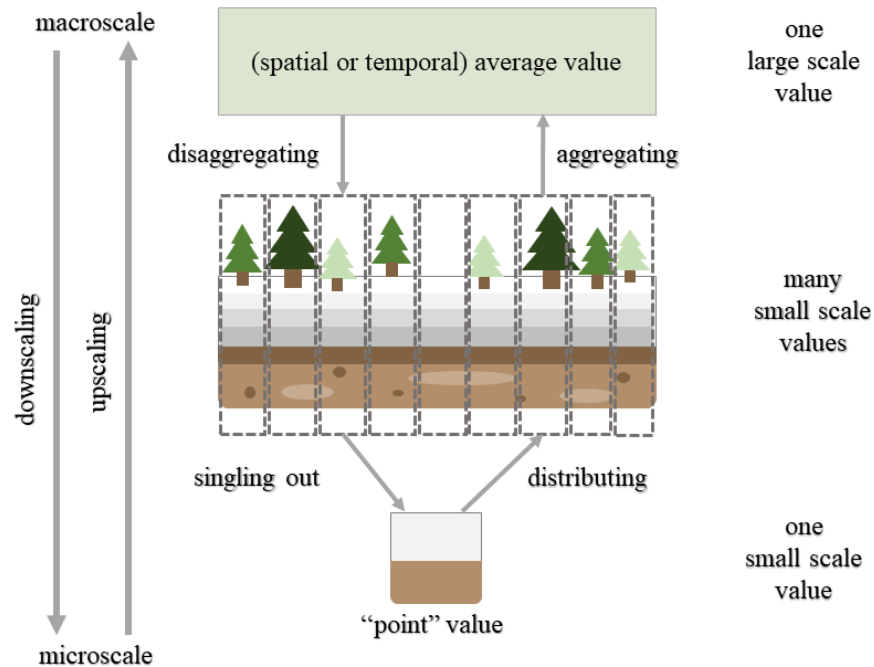


Figure 2.1: Conceptual illustration of upscaling and downscaling processes, inspired by Blöschl and Sivapalan (1995).

2.3 Upscaling in Hydrologic Modelling

Upscaling refers to the process of translating heterogeneous small-scale information into a representative form, either statistically or parametrically, at a larger, homogeneous scale. This is typically achieved by deriving a functional relationship for a large-scale constitutive system using information about processes and heterogeneity at the small scale (Blöschl and Sivapalan, 1995). Upscaling follows a two-step process of distribution and aggregation. First, microscale information is distributed across a catchment area, generating a spatial distribution of the relevant parameter, state, or flux. This distributed data is then aggregated into a single value or a small set of values, such as a mean and standard deviation, to represent the parameter at the macroscale (Blöschl and Sivapalan, 1995). This process inherently simplifies the system, leading to loss of small-scale heterogeneity information. Different upscaling strategies employ distinct approaches to address this loss of subgrid information as grid cell sizes increase. There are two primary approaches to upscaling: bottom-up and top-down. In the bottom-up

approach, large-scale relationships are derived by explicitly translating small-scale heterogeneity into a useful large-scale representation. This method includes techniques such as exact averaging and effective parameterization. Effective parameterization is the application of known small-scale constitutive relations at a larger scale using modified effective parameter values to reproduce the large-scale flux (Ranjram, 2021). In contrast, the top-down approach infers large-scale relationships from observed aggregate system behavior, such as streamflow, using methods like empirical analysis and inverse modelling to determine how to describe the larger-scale outcome without information about local detail.

Bottom-up upscaling methods can be further categorized into four classes: spatial averaging, stochastic methods, effective parameterization, and bulk descriptions (Ranjram, 2021). Each approach addresses subgrid variability in different ways while aiming to incorporate data collected at smaller scales and account for the influence of unresolved heterogeneities in the model (Blöschl and Sivapalan, 1995; Grayson and Blöschl, 2000; Harter and Hopmans, 2004). Spatial averaging is a basic form of upscaling, where a distribution of fluxes or states is condensed into a single mean value, disregarding other local spatial details. Stochastic approaches represent subgrid variability using probability density functions (PDFs), which capture the distribution of subgrid parameter values within the computational element without record to their spatial pattern. The upscaled value of interest is generated by aggregating these distributed values using ensemble methods (e.g., Beven and Kirkby, 1979; Craig et al., 2010; Moore, 1985). For example, Li and Pomeroy (1997) used PDFs of blowing snow depth to develop a potential scaling function for upscaling fluxes from point to area. However, a key drawback of is their computational complexity when conducting fine-scale simulations over large geographical areas, as multiple observations are needed to distinguish actual signals from random variability (Clark et al., 2017; Fatichi et al., 2016). Effective parameters are modified values used in model equations to approximate the aggregate behaviour of a system by treating subgrid heterogeneity through the substitution of a single macro-scale parameter into known relations derived at the micro-scale in place of micro-scale parameters (Neuman and Di Federico, 1998). Finally, bulk descriptions for upscaling replace subgrid variability with a single lumped equation, such as using hydraulic conductivity in Darcy’s Law instead of detailed geometric information to represent complex subgrid characteristics like a soil’s pores network. While both approaches simplify sub grid complexity, effective parameterizations retain the form of micro-scale equations with modified parameters, whereas bulk descriptions may alter the equation’s structure entirely to represent aggregate behaviour, often using empirical relationships.

2.4 Constitutive Relationships

Constitutive relationships characterize the interactions between system states (e.g., snow cover area) and fluxes (e.g., snowmelt) in a model. These relationships are typically observed and formalized at small scales, where observational methods can effectively capture them. For instance, constitutive relationships for snow fluxes are derived by integrating physical laws and empirical data, using mass and

energy balance equations combined with field measurements to describe processes such as snow melting, sublimation, and compaction under varying environmental conditions (Han et al., 2023; Mishra and Mahajan, 2004). However, observing and quantifying fluxes at the watershed scale is challenging due to technological limitations, making it difficult to derive watershed-scale constitutive equations directly, instead, upscaling techniques are needed.

Constitutive relationships at the watershed scale can be estimated using results from high-resolution models. This type of upscaling process simplifies relationships between model inputs and outputs by synthesizing results from fine-scale models, which are then aggregated to a coarser resolution. Snowden and Craig (2016) applied this approach to upscale groundwater-surface water exchange fluxes from the MODFLOW model, revealing a power-law relationship between average storage and flux. Ranjram (2021) explored the relationship between signature parameters and aggregate descriptions of hillslope parameters that control hillslope-scale subsurface flow, demonstrating how constitutive relationships can be derived by interrogating the results of fine-resolution models. This approach will be applied in this thesis.

ML offers new opportunities for identifying constitutive relationships in hydrology. Tartakovsky et al. (2020) employed physics-informed deep neural networks to learn constitutive relationships in subsurface flow problems, by integrating physical constraints like Darcy’s law and Richard’s equations. This approach allows ML models to infer relationships constrained by physical principles, enhancing parameter estimation accuracy without relying solely on direct measurements. In this thesis, ML methods constrained by physically based models will be used to infer constitutive relationships describing snow-related processes, avoiding direct integration of equations or purely regression-based approaches. This method leverages the strengths of ML while maintaining a strong foundation in the physical principles of hydrology, offering a novel pathway for improving the representation of snow fluxes at larger scales.

2.5 Downscaling in Hydrologic Modelling

Downscaling is the process of transforming coarse-scale information into finer-scale information (Li et al., 2016). The hydrometeorological literature typically reports two primary downscaling methods: dynamical and statistical. Dynamical downscaling refers to the use of the finer scale regional climate models (RCMs) that apply atmospheric physics over a region, utilizing boundary conditions from coarser Global Climate Models (GCMs) to produce high-resolution outputs. In contrast, statistical downscaling relies on observed data to derive relationships between large-scale atmospheric variables, obtained either from climate model outputs or observations, and local-scale climate variables (Wilby and Wigley, 2000). Statistical approaches often involve spatial disaggregation and bias correction, using methods such as interpolation (Eum and Cannon, 2017) or climate analogs to replicate small-scale climate observations

(Abatzoglou and Brown, 2012; Hidalgo et al., 2008; Hwang and Graham, 2013). Here, the focus is on statistical downscaling methods, since dynamical downscaling is not relevant to this study.

A diverse range of statistical downscaling techniques has been developed in recent years, generally falling into one of three major categories: stochastic weather generators, weather typing schemes, and regression (transfer function) methods (Wilby et al., 1998; Nguyen et al., 2006; Vrac and Naveau, 2007). Stochastic weather generators (Wilks and Wilby, 1999; Olsson et al., 2009) are statistical models that generate local-scale time series based on probability density functions whose parameters are related to large-scale data (e.g., Vrac and Naveau, 2007). Weather typing, on the other hand, classifies large-scale atmospheric patterns into distinct weather types, which are then linked to local-scale meteorological conditions to improve climate model predictions at finer spatial resolutions (Fowler et al., 2007). Regression-based downscaling methods establish a direct quantitative relationship between local variables and large-scale information using regression functions (Bürger and Chen, 2005). These methods vary based on the transfer function, predictor variables, and statistical fitting procedures used, with popular options including multiple linear regression and artificial neural networks. In hydrologic models, regression-based downscaling is typically applied to forcing variables such as precipitation, temperature, and wind, downscaling them from weather models to match the finer model scale of the hydrologic model. For instance, Wang et al. (2020) used recurrent neural networks to downscale temperature and precipitation, which were then incorporated into the Soil and Water Assessment Tool (SWAT) model to simulate processes such as flow and evapotranspiration (ET).

A specific subset of statistical downscaling, frequently used in hydrologic models to adjust GCM outputs, is the application of lapse rates. Lapse rates refer to the rate of temperature decrease with height and are particularly useful in mountainous terrain, where elevation is one of the primary controls on precipitation and temperature. The characterization of lapse rates has several applications, particularly in downscaling reanalysis data and weather predictions in complex terrain. By empirically estimating lapse rates, statistical relationships between elevation and climate variables, from high-resolution topographic and climatic data, these rates can be used to adjust coarse resolution climate model outputs based on fine scale elevation (Praskievicz, 2018; Sheridan et al., 2010). For example, Dutra et al. (2020) applied lapse rates to downscale ERA5 temperature and snow-related variables in complex terrains. Similarly, Praskievicz (2018) used local topographic lapse rates to downscale maximum and minimum temperatures and precipitation from GCMs to finer spatial resolutions in mountainous regions. Like other statistical downscaling techniques, the lapse-rate method relies on empirical relationships between variables such as elevation and climate. However, it also shares similarities with dynamic downscaling, as it effectively enhances the grid resolution of climate model outputs by incorporating finer-resolution topographic data.

2.6 Machine Learning in Hydrology

The application of ML methods has advanced rapidly in hydrology, driven by improvements in computational power and the availability of large datasets. ML has been applied to various aspects of hydrology, including streamflow forecasting (Ng et al., 2023; Troin et al., 2021) and water quality assessment (Ahmed et al., 2019). Unlike traditional statistical methods, which rely on predefined assumptions about the relationships between model inputs and outputs, ML does not require assumptions about the data model structure. Instead, ML methods seek to detect and describe arbitrary relationships between multivariate hydrological datasets, generalize them, and make predictions. This ability allows ML to handle large datasets effectively and identify patterns in hydrological variables, which are often highly nonlinear due to the inherent heterogeneity of hydrologic systems. Consequently, ML is a powerful tool for advancing water resource management.

The process of learning functional relationships between data can be broadly classified into supervised and unsupervised learning approaches. The key distinction between the two lies in the use of labeled data that defines input and output datasets. Unsupervised learning analyzes and clusters unlabeled data, discovering hidden patterns without human intervention. In contrast, supervised learning methods train algorithms to classify data or predict outcomes accurately by learning from labeled datasets. Once trained, the model is tested on a separate independent dataset to evaluate its accuracy (Lange and Sippel, 2020). Supervised ML algorithms can be further categorized by the algorithms used into neural network-based, support-vector-based, and tree-based methods. In this thesis, supervised methods are used to derive upscaled constitutive relationships, while unsupervised techniques, such as clustering, are employed to identify similar groupings of HRUs.

Neural networks are among the most widely used ML algorithms in hydrology (Shen, 2021). Artificial Neural Networks (ANNs) have been particularly successful in modelling nonlinear behaviors such as rainfall-runoff (Gholami and Sahour, 2022; Mao et al., 2021), streamflow (Hunt et al., 2022), water quality (Wang et al., 2022), and groundwater systems (Wunsch et al., 2021). Inspired by the human brain, ANNs consist of interconnected units called neurons. A common type of ANN is the multilayer perceptron, which is a basic feedforward ANN that processes input data through multiple layers of neurons. As shown in Figure 2.2, multilayer perceptron consists of an input layer, hidden layers, and an output layer. Inputs are fed into the input layer, which processes the data through multiple layers of interconnected neurons. Each neuron computes a weighted sum of its inputs, adds a bias, and applies a nonlinear activation function. The network learns by adjusting these weights and biases during training, allowing it to capture complex patterns and make predictions based on the relationships it learns from the data (Xu & Liang., 2021). The final result is produced in the output layer after the processed information passes through the hidden layers. However, one of the major challenges with ANNs is their lack of transparency. The processes of these models are often considered opaque, making it challenging for hydrologists to fully understand the underlying relationships (Barakat and Bradley, 2010).

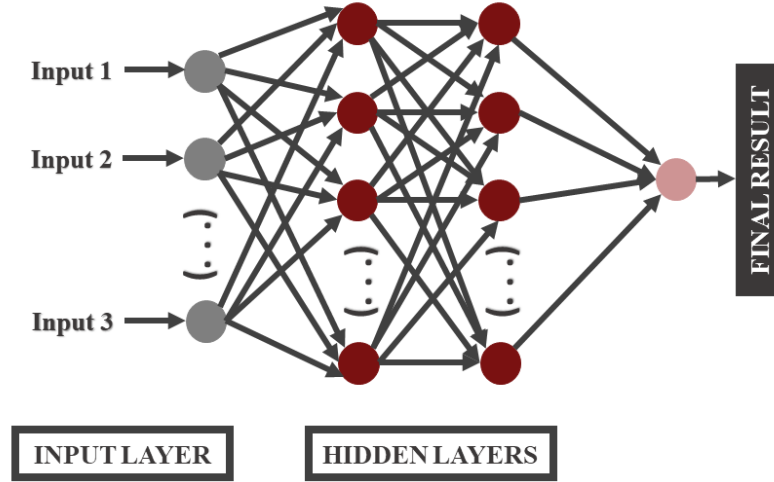


Figure 2.2: Schematic of a basic feedforward multilayer perceptron architecture, where the input values (grey circles) pass through multiple hidden layers, where each neuron processes the inputs with weights, biases, and activation functions. The final result (pink circle) is produced in the output layer after processing, with the network adjusting weights and biases to minimize error and improve predictions.

To address the challenge of poor interpretability, interpretable ML methods have gained increasing attention. Interpretability refers to a model's transparency and the clarity of its internal logic, enabling users to understand how inputs are transformed into outputs (Doran et al., 2017; Guidotti et al., 2018). Tree-based algorithms, such as Random Forests (RF) and eXtreme Gradient Boosting (XGBoost), are often considered relatively interpretable because they allow for the evaluation of feature importance, helping to assess the influence of specific input variables on model predictions. Additionally, these models can visually represent decision-making processes through graphical depictions of decision trees, further enhancing transparency (Chakraborty et al., 2021).

Tree-based algorithms, such as Decision Trees (DT), XGBoost, and RF, work by recursively splitting data into subsets based on threshold values, a process known as recursive partitioning. In DTs, this process continues until further splits no longer improve model performance, after which insignificant branches are pruned to simplify the structure (Leiva et al., 2019; Rokach and Maimon, 2005). XGBoost, an ensemble method based on gradient-boosted DT, combines predictions from multiple "weak" learners to create a "strong learner" through an iterative training process. Variations of XGBoost models have been successfully used to estimate water levels in hydrologic studies, particularly interpretable ML studies (e.g., Wang et al., 2024a; Xu et al., 2024). However, their performance largely depends on selecting the right variables and can be prone to overfitting (Han et al., 2019). XGBoost models are also known to be "data-hungry," often requiring large amounts of data to perform well, with lower-quality predictions when insufficient data is available. In contrast, RF models, while sensitive to input variable selection, have strong resistance to overfitting (Lange and Sippel, 2020).

The RF model, proposed by Breiman (2001), is an ensemble ML technique that combines classification and regression tree methods with statistical learning theory. It consists of multiple decision trees generated through a bootstrap sample of the training data and random variable subset selection (Pang et al., 2017), as shown in Figure 2.3. The final prediction is produced by aggregating the outputs from these trees, by averaging (for regression tasks). RF models have been successfully applied across various fields, including groundwater studies (Baudron et al., 2013), flood hazard assessment (Wang et al., 2015), remote sensing (Zhao et al., 2018b), downscaling (Pang et al., 2017), and landcover classification (Tian et al., 2006). One of the key advantages of RF models is their ability to handle large datasets with correlated variables, thanks to their nonparametric nature, which enables them to manage multivariable and nonlinear problems (Breiman, 2001).

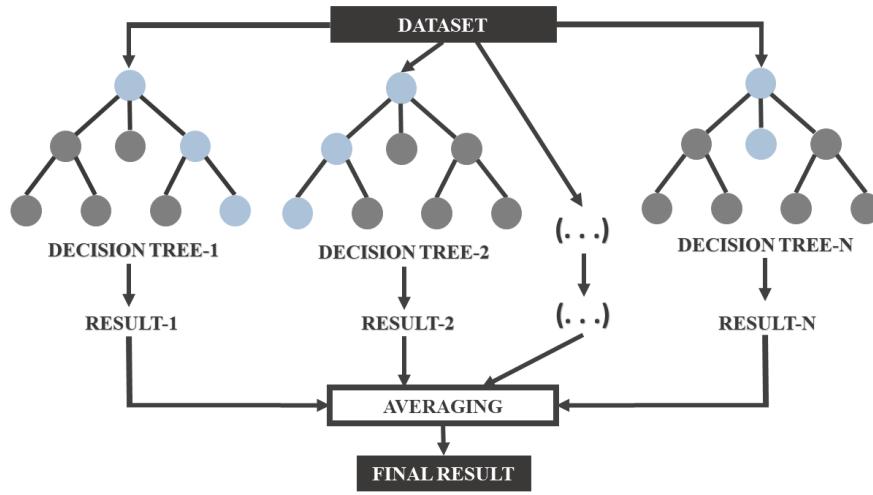


Figure 2.3: The random forest model selects features and builds a forest of decision trees by splitting the initial values into subsets (represented by grey circles) for each tree. The mean squared error is then calculated independently for each subset, with the blue circles representing the values that result in the lowest mean squared error.

There are additional tools to aid in the interpretability of ML algorithms by revealing the factors driving uncertainty and influencing predictions. For instance, the Gini feature importance index, commonly used in tree-based models, measures the reduction in Gini impurity—a metric of data purity—when a feature is used to split data. The greater the impurity reduction across the tree, the more critical the feature, providing insight into which variables most strongly influence model decisions (Gregorutti et al., 2017; Menze et al., 2009). Similarly, SHapley Additive exPlanations (SHAP), grounded in cooperative game theory, assign each feature a SHAP value that represents its contribution to a specific prediction (Aydin and Iban, 2023). Unlike the Gini index, SHAP not only highlights feature importance but also indicates whether a feature has a positive or negative impact on the prediction and to what extent. These methods are essential for making complex models more transparent and interpretable.

Physics-informed ML refers to a class of models that integrate fundamental physical laws and knowledge of the domain in ML models to establish predictive models that are both scientifically robust and physically consistent (Kashinath et al., 2021). Recent integration of physics-informed ML models with hydrological models aims (Deng et al., 2024) to address the lack of explicit interpretability by combining the strengths of ML models, which excel in advanced data mining, and hydrologic models, which incorporate physical restraints of hydrologic processes. Bhasme et al. (2022) developed a physics-informed ML model that integrates the conceptual ABCD hydrological model with ML algorithms. This approach enhances predictive accuracy by using the ABCD model to maintain physical consistency while ML identifies complex relationships between inputs and outputs, such as ET and streamflow. In some cases, ML has been constrained by the physical outputs of hydrologic models, as demonstrated by Jiang et al. (2020), where a deep learning model improved streamflow predictions by learning from simulation errors in process-based models. This is conceptually similar to the nudging of tendencies approach, where model outputs are iteratively adjusted toward the observed behaviour. This approach allows deep learning models to retain physical realism while benefiting from data-driven insights.

Long Short-Term Memory (LSTM) networks, a type of recurrent neural network, are designed to capture long-term dependencies in sequential data by using memory cells and gates to regulate information flow, enabling dynamic modelling of temporal relationships (Hochreiter and Schmidhuber, 1997; Sahoo et al., 2019). This ML algorithm has gained significant popularity in hydrology due to its ability to dynamically learn temporal relationships, making it well-suited for time series forecasting such as continuous rainfall-runoff modelling (Kratzert et al., 2018) or short-term streamflow forecasting (Arsenault et al., 2023). In the context of physics-informed ML, LSTMs have been combined with physical constraints or hybridized with physics-based models to ensure consistency with fundamental principles, such as mass conservation (Bhasme et al., 2022). For example, Lu et al. (2021) developed a hybrid LSTM model that integrates the outputs of physics-based hydrological model as additional inputs, enhancing predictions in data-scarce basins by leveraging both data-driven insights and physical knowledge. Despite their popularity and proven success in time-series forecasting, LSTMs are inherently designed to process sequential data, making them unsuitable for studies that focus on static datasets or spatial modelling tasks where temporal dependencies are not a concern (Sabzipour et al., 2023).

2.7 Scaling with Machine Learning

ML-based methods have increasingly been used in scaling applications, particularly for downscaling, in environmental and geospatial sciences. These models are typically trained using large-scale reanalysis or remote sensing datasets and corresponding local-scale observations to estimate finer-scale local variables from coarse-resolution inputs. For example, Ke et al. (2016) applied RF, Cubist, and Support Vector Regression (SVR) to downscale MODIS 1 km ET data to 30 m resolution using Landsat 8-derived indices as predictors. The models were trained using upscaled Landsat data at 1 km resolution to

match MODIS ET values and then applied to finer-resolution Landsat data to generate high-resolution 30 m ET estimates. Yatheendradas and Kumar (2022) developed a deep learning model to downscale MODIS (Moderate Resolution Imaging Spectroradiometer) snow cover fraction (SCF) from 5 km to 1 km using a convolutional neural network trained on partially cloud-free MODIS scenes, with topographic and land surface variables like snow albedo and land surface temperature to improve predictions in gap-filled regions. Similarly, Baird (2025) applied a RF model to downscale ERA5-Land snow depth (9 km) to 50 m resolution using topographic variables and Landsat-derived snow cover frequency, training the model on high-resolution lidar snow depth observations from multiple basins in Colorado. These studies collectively demonstrate that ML-based methods can produce reliable downscaling results. These ML methods are effectively downscaling remotely sensed data products, rather than model outputs, and are therefore unable to downscale variables which are not directly observed, like melt or sublimation fluxes.

ML-based methods also show significant potential for upscaling. Ge et al. (2019) highlighted that ML models can produce more accurate upscaled results than traditional regression and geostatistical approaches. Unlike methods based on single or multiple site measurements, ML-based upscaling is not limited to areas with available in situ data, enabling continuous spatial application. Moreover, ML models excel at handling complex and nonlinear relationships. For instance, Wang et al. (2021) modeled the relationship between in situ albedo measurements and land surface variables (such as NDVI, temperature, elevation, and landcover) using ML models like RF, K-Nearest Neighbors, and Cubist. These models, trained on point-scale data, were applied to larger spatial scales (500 m and 1 km), with RF and K-Nearest Neighbors proving particularly effective in upscaling in situ albedo measurements. Yin et al. (2022) trained an RF model to upscale GRACE-based groundwater storage estimates by establishing a relationship between coarse-resolution satellite data (1°) and high-resolution input variables such as ET and soil moisture. After training on coarse data, the RF model predicted finer-resolution water storage (0.25°) using these inputs. The model's performance was further enhanced through error correction using Kriging interpolation, resulting in improved accuracy when validated against in-situ groundwater measurements.

While ML is often perceived as a "black box," where data inputs lead to decisions without clear visibility into the intermediate processes, this perception is more a reflection of limited inspection than an inherent limitation of the models themselves (Nearing et al., 2021). In hydrology, cross-validation techniques and diagnostic tools are crucial for validating models and extracting insights, and these can also be applied to ML models to better understand the contributions of inputs. Methods such as sensitivity analyses can enhance our ability to interpret the information within these models and assist in validating interpretable ML methods (Sundararajan et al., 2017). Although ML-based upscaling has been used in hydrology, little attention has been given to designing ML architectures that incorporate scale-dependent physical process constraints. Moghaddam et al. (2021) explored the ability of ML algorithms to infer effective hydraulic conductivity from binary hydraulic conductivity grids. While the algorithm extracted relevant energy dissipation information, the results were not immediately interpretable and

required additional steps to convert the data into a human-readable format. Aside from this study, few have explored the potential of ML to infer hydrological upscaling rules, a gap that this thesis aims to address.

2.8 Snow Processes

Snow plays a critical role in water supply for cold regions, particularly in mountainous areas, where much of the precipitation falls as snow, and the majority of the annual streamflow is driven by snowmelt (Gray and Male, 1981; Doesken and Judson, 1997). In these areas, accurately simulating key snow processes—such as accumulation, sublimation, and melt—and tracking snow properties like depth, density, and water equivalent is essential (Lv et al., 2019). However, the spatial distribution of snow is often uneven, and in-situ measurements are limited due to challenges like low population density, harsh environments, budget constraints, and the extensive post-processing required to generate reliable data (Fassnacht, 2004; Sextstone et al., 2016). Therefore, advancing our understanding of snowpack evolution depends, in part, on improving knowledge of point-scale processes and their associated spatial variability.

State and flux variables are critical to understanding snowpack dynamics. State variables quantify the amount, phase, or distribution of water in a mass balance model. For instance, the snow water equivalent (SWE) measures the amount of water present in a snowpack, expressed as the depth of water if the snowpack were completely melted. Snow-covered area (SCA) indicates the fraction of the landscape covered by snow, ranging from 0 to 1. Flux variables represent the rates of change of a mass balance, describing the movement of water mass into and out of storage elements per square meter of land surface. Key fluxes include snowmelt, the process by which snow transitions into liquid water, and sublimation, the direct conversion of snow or ice into water vapor without passing through a liquid phase. A simplified mass balance on SWE [mm] within a computational element can be expressed as:

$$\frac{d\overline{SWE}}{dt} = \overline{S}_{snowfall} - \overline{S}_{sublim} - \overline{S}_{melt} \pm \overline{S}_{blow} \quad (2.1)$$

here $\overline{S}_{snowfall}$ [mm/d] is the (spatial) mean accumulation of snow at the surface, $\overline{S}_{sublim}$ [mm/d] is the mean surface sublimation rate, $\overline{S}_{melt}$ [mm/d] is the mean divergence of water after melt and runoff, and $\overline{S}_{blow}$ [mm/d] is the mean addition or removal of snow via blowing snow.

Snowmelt begins when net energy inputs from atmospheric and radiative forces become consistently positive and the snow heats up to 0°C, reaching a threshold that triggers the conversion of snow into liquid water with further energy input (Dingman, 2015). Shortwave radiation is a driving force for snowmelt as it provides direct solar energy that heats the snow surface. Air temperature also plays a crucial role, as warmer temperatures increase the energy available for melting via convection at the snowpack surface. Snow loss through sublimation results in water loss from the snowpack to the

atmosphere (Stigter et al., 2018), and is influenced by energy availability, wind speed, and relative humidity. Higher wind speeds promote sublimation by increasing turbulent diffusion of snow particles, while higher humidity tends to reduce sublimation rates as supersaturated air promotes deposition (Déry et al., 1998). Blowing snow occurs when wind entrains loose surface snow particles. It involves two processes: the horizontal transport of snow by wind and the sublimation of blowing snow. Like sublimation, blowing snow is influenced by wind speed and humidity.

2.8.1 Upscaling Snow Fluxes

Snowmelt is often implicitly upscaled using Snow Depletion Curves (SCDs), which graphically represent the relationship between SCA and mean SWE in an area. The shape of the depletion curve is implicitly related to the frequency distribution of SWE. Under homogeneous melt conditions, the SCD can be described by relating SCA to a shift in the PDF of accumulated SWE at a known initial state (Pomeroy et al., 1998; Liston, 1999; Luce et al., 1999; Essery & Pomeroy, 2004). Luce et al. (1999) introduced a dimensionless depletion curve from the PDF:

$$SCA = \int_{SWE_{acc}}^{\infty} f_{SWE}(SWE) \quad (2.2)$$

where SWE_{acc} is the accumulated SWE melted since peak accumulation, and f_{SWE} denotes the initial probability distribution of SWE. In this equation, SWE is the variable of integration, and the integral computes the fraction of the area where snow remains by evaluating the probability of SWE exceeding the accumulated melt. The SCD accounts for snow variability by characterizing the relationship between SCA and SWE, with its shape determined by the initial SWE distribution across the basin. Common distributions for representing initial SWE include lognormal, gamma, and normal distributions (Donald et al., 1995; Liston, 2004; Godio and Rege, 2016). Different distributions yield varying levels of fit to field observations. In high-elevation basins of Alberta, complex terrain and wind redistribution produce highly variable snowpacks, resulting in broader SWE distributions and more gradual, concave SCDs. Conversely, in low-elevation regions such as Southern Ontario, where snow accumulation is typically more uniform and terrain less variable, the SWE distribution is narrower, and the SCD tends to be steeper and more linear, reflecting a more abrupt transition from snow-covered to snow-free conditions.

Upscaling methodologies that offer a more sophisticated form of aggregation can complement or replace the use of SCDs, particularly in complex terrains. Gupta et al. (2023) aggregated variables such as snow depth, snow cover, shortwave radiation, and wind speed by distributing meteorological forcing across the catchment based on topographic features like slope, aspect, and curvature. The model first simulated these variables at a hyper-resolution (10 m), capturing fine-scale processes, and then spatially averaged or aggregated them over the larger catchment area to integrate these values into a hydrological model (ParFlow-CLM) to analyze basin-wide snowmelt, runoff, and ET. Similarly, Pomeroy (2004) aggregated SWE and melt data across different landscape classes (boreal forest, shrub tundra, and alpine

tundra) in a subarctic mountain catchment. This study used spatial statistics, such as log-normal and bimodal distributions, to represent SWE frequency and employed covariance between SWE and melt rate to account for spatial variability in snow depletion. However, these upscaling methods rely heavily on assumptions about the representativeness of spatial averages and statistical distributions, which may not fully capture non-linear interactions or threshold behaviors at finer scales.

2.8.2 Downscaling Snow

Downscaling in the context of snow simulation typically involves downscaling forcing variables, which are provided at coarse resolutions, to the resolution of hydrologic models to simulate snow processes at finer spatial resolutions (Thornton et al., 2021; Mital et al., 2022). Statistical and dynamic downscaling methods have been developed to reconstruct meteorological variables to finer scales by adjusting meteorological variables according to known topographical effects (Bernhardt et al., 2010; Helbig et al., 2024), interpolation (Pflug et al., 2021), and ML (King et al., 2020). Generally, these downscaled forcing variables are first adjusted to match the spatial scale of interest, and subsequently integrated into a hydrologic model to simulate related snow state and flux variables, such as SWE, snowmelt, and sublimation. For example, Berg et al. (2024) evaluates the effectiveness of downscaled dynamically downscaled meteorological forcing data from the High-resolution Intermediate Complexity Atmospheric Research model combined with a snow transport model (FSM2oshd) to downscale spatial snow accumulation patterns at resolutions as fine as 50 meters in alpine terrains to model snow transport processes in complex topography. Iseri et al. (2021) used a dynamic downscaling methodology to downscale coarse-resolution global reanalysis atmospheric forcing datasets to finer 3 km spatial resolution. The downscaled atmospheric data was then used as an input to the Watershed Environmental Hydrology Hydro-Climate Model (WEHY-HCM) to simulate fine-scale snow accumulation and melting using an energy balance approach.

Snow cover fraction is related to downscaling, as it represents the proportion of a grid cell covered by snow (a sub-grid distribution property). Snow cover fraction directly impacts energy balance calculations, acting as a modifier for (e.g.,) snowmelt because only the SCA can melt. In other words, when only a portion of the surface is snow-covered, energy inputs like solar radiation contribute to melting only that fraction, limiting total snowmelt to the extent of the snow-covered ground. Li et al. (2015) implemented a snow cover fraction downscaling method that utilized topographic factors such as slope, aspect, and solar radiation to model snow ablation at finer resolutions, achieving more detailed snow distributions in complex terrains. These approaches highlight the importance of considering topographic influences to effectively capture system heterogeneity when downscaling meteorological and snow-related variables.

Numerous studies have demonstrated the potential of combining remote sensing products with reanalysis or model-based datasets to produce high-resolution SWE estimates. Shao et al. (2022) used a regularized linear regression approach to integrate multiple coarse-resolution SWE products, such as

ERA-Interim, GLDAS, and AMSR-E/AMSR2, into a high-resolution dataset for the Northern Hemisphere. Similarly, Mital et al. (2022) developed a RF model trained on lidar-derived SWE observations and gridded meteorological variables to estimate SWE at unmapped time points in mountainous regions. These approaches highlight the growing use of ML to improve SWE estimation from remote sensing and large-scale datasets. However, despite these advancements in downscaling forcing variables and remotely sensed datasets, applications of directly downscaling hydrologic model outputs of snow-related state and flux variables appear to be limited or not well documented in the literature.

Chapter 3

Upscaling snow accumulation and ablation using machine learning

3.1 Introduction

The ability of hydrological models to represent the world is constrained (in part) by the large degree of heterogeneity in landscape properties and the complexity of watershed responses to spatiotemporally varying climatic inputs (Troch et al., 2009). Heterogeneity can lead to significant spatial variation in landscape water storage and fluxes, due to variations in regional topography, climate, soil properties, landscapes, and land-use practices (Cheng et al., 2014; Yang et al., 2019). The need for computational scaling methods arises directly from this heterogeneity of the natural world (Klemeš, 1983), as it is typically infeasible to simulate all details of a heterogeneous watershed due to both computational cost and a lack of observational data. This challenge is particularly pronounced in mountainous regions, where complex terrains, with variations in elevation, slope and aspect, results in a high degree of heterogeneity. Moreover, observational data in these areas is sparse, and significant portions of the terrain lie at elevations higher than most meteorological gauges (Lundquist et al., 2019; Wayand et al., 2013). These limitations can lead to an incomplete (or inadequate) representation of hydrological states and fluxes at scales of interest.

Hydrological models use constitutive relations which describe the relationship between system states (often water storage) and water fluxes at the computational scale, which is often larger than the representative scale of heterogeneity. Given an initial state of the system, hydrological models apply these constitutive relations to calculate fluxes (e.g., runoff, infiltration, or snowmelt) and attempt to describe the movement of water through a catchment and update the system state. However, a fundamental issue

in the application of these constitutive relations is one of scale. For instance, many constitutive relations are derived for small-scale homogeneous systems, where the physics of the hydrological process is reasonably well-characterized and then applied at a much larger scale, where heterogeneity is influential. At these larger scales, the fine-scale constitutive relations instead become useful approximations (Beven and Germann, 2013). These approximations typically require the use of effective parameters, values that are often estimated through calibration, where model outputs are tuned to match observations by adjusting input parameters. In this context, calibration compensates for unresolved subgrid variability by modifying parameter values to reflect the aggregate behavior at larger scales, rather than their true physical meaning at fine scales, thus enabling adequate predictions across scales. Hydrological modelling tries to alleviate this problem through the discretization of a watershed into scale-appropriate, relatively homogenous units based on landscape properties, referred to as HRUs (Kampf and Burges, 2007). However, it is still challenging to capture the influence landscape heterogeneity plays within these HRUs, particularly if they are relatively large in size. Neglecting this sub-HRU spatial heterogeneity potentially compromises the reliability of model forecasts and predictions and makes it more difficult to parameterize models without calibration. In addition, the implicit neglect of heterogeneity makes it difficult to ascertain its role in influencing the hydrologic response of the landscape. Lastly, the ability to properly represent the influence of sub-HRU scale heterogeneity enables the use of larger HRUs, leading to more computationally efficient models. Future hydrological models may, therefore, benefit from the identification of scaling approaches in complex landscapes that can both provide insight and improve our ability to simulate complex heterogeneous watersheds.

Upscaling refers to a change in computational resolution from a smaller to a larger scale (Gupta et al., 1999), often in the form of a mapping from a heterogeneous system to an equivalent homogeneous one at a larger scale. Thus, a computational upscaling method is a set of rules that performs a many-to-one conversion, where the many heterogeneous fluxes within a domain (resulting from heterogeneity in parameters and states) are converted into a single mean flux that adequately represents the influence of small-scale heterogeneity. Computational upscaling strategies are useful when governing constitutive relations are to be applied at a scale larger than the scale they were derived, or when a new relationship is derived to express the net impact of heterogeneous small-scale fluxes at a larger scale, which otherwise would require brute force computation or unavailable data (Blöschl and Sivapalan, 1995). There is a rich history of using a variety of methods to identify upscaling relations in hydrology. These include methods focused on intelligent approaches for estimating effective parameters (e.g., Feddes et al., 1993; Samaniego et al., 2010), the use of distribution-based methods such as those of (e.g.,) Wood et al. (1986) and Zhao (1992) and the development and solution of stochastic differential equations that explicitly track probability distributions of sub-grid states (e.g., Kavvas, 2003; Yoon and Kavvas, 2003; Ohara et al., 2008, 2014). This study focuses on developing novel techniques to detect upscaling relationships through the application of ML methods informed by fine-resolution hydrological models.

With the ever-increasing availability of both computational resources and data, data-driven approaches like ML, are increasingly used in the field of hydrology. ML models are now regularly (and successfully) applied to water quantity and quality forecasting (Ghobadi and Kang, 2023; Sit et al., 2020; Wang et al., 2022; Varadharajan et al., 2022). Most of these models look to either replace or augment physical/conceptual hydrological models by transforming input signals (such as precipitation and temperature) into output signals (streamflow), with mixed degrees of regard for the internal function of the watershed. For instance, comparative studies involving traditional statistical models and various ML models, such as LSTM models, have successfully shown the ML model's ability to outperform water-level predictions in hydrological studies (Ahmed et al., 2022; Herath et al., 2023). However, despite recent advances in their predictive performance and ability to handle highly complex spatiotemporal datasets, the application of ML algorithms in hydrological upscaling remains nascent. The limited application of ML algorithms in upscaling applications is in part due to their limited interpretability (Gilpin et al., 2018; Ancona et al., 2017; Schmidt et al., 2020), which stems from characterization of ML methods as 'black-box': data goes in, decisions come out, but the processes between the input and output are often opaque. As a result, studies often use ML-generated mappings from measured inputs to outputs without consideration of the underlying physical processes involved (Yosinski et al., 2014).

To date, little consideration has been given to the use of ML as an alternative means of identifying how watershed fluxes can be internally calculated within a model (i.e., constitutive relations). A select number of studies (e.g., Bennett and Nijssen, 2021; Moghaddam et al., 2021; Yuval and O'Gorman, 2020), have successfully explored ML algorithm's ability to derive and incorporate process-based relationships into existing hydrological models. For instance, Moghaddam et al. (2021) examines the ability of ML algorithms to infer the effective hydraulic conductivity of binary hydraulic conductivity grids. Bennett and Nijssen (2021) derive deep learning-based parameterizations of turbulent heat fluxes and integrate them into the SUMMA model (Clark et al., 2015). Although this deep learning process has been used to represent individual hydrologic process-based relationships, it has not yet been explicitly applied to address upscaling in hydrology. The application of ML algorithms, specifically RF models, in hydrologic upscaling has been limited to the interpolation of point-scale measurements rather than the generation of scaling relations for estimating water fluxes. For instance, Xu et al. (2018) employ five ML algorithms to extrapolate the spatial distribution of ET from flux towers at the point scale to the regional scale as a form of upscaling. However, the study describes an interpolation methodology generating multiple values across the basin, rather than upscaling the small-scale information to a bulk representation, and therefore deviates from the definition of computational upscaling we use here. In contrast, Yu et al. (2024) introduced an XGBoost-based upscaling approach that utilized point-scale soil moisture measurements to train the XGBoost algorithm. The model learned correlations with geospatial features such as albedo, Normalized Difference Vegetation Index (NDVI), and land surface temperature, enabling it to extrapolate soil moisture measurements across a larger area of 100 km × 100 km. In this paper, we will examine how

these ML algorithms can first emulate the large-scale behaviour of heterogeneous hydrological systems, and thereafter be used to infer or replace such scaling laws.

This study focuses on developing and testing a general upscaling methodology to derive constitutive relations via ML techniques trained on fine-resolution model output. A case study focussing on the simulation of snowpack evolution in a mountainous area provides a basis to evaluate the framework of steps and assess the ML upscaling methodology. The objectives of this study are to (1) develop, apply, and test an ML-based upscaling methodology to examine and characterize relationships between states, fluxes, and forcings across the land surface, (2) evaluate the predictive ability and interpretability of an ML algorithm to identify whether the algorithm provides an accurate simulation for mean landscape fluxes, and (3) apply the ML-estimated constitutive relations between systemwide storage and fluxes to relatively larger scales (5-20km²) to simulate coarse-resolution mean hydrologic system response with low-order statistics of state and without the need for fine-scale (~0.05km²) simulations.

The case study demonstrates the ML upscaling methodology is skillful at simulating the interactions between the atmosphere and local terrain features that influence the local-scale variability of snowmelt and sublimation. Although we limit ourselves here to a model of only snowpack to limit the number of confounding factors, the approach may, in principle, be applied to more complex hydrological models. Snow ablation may be considered a relatively isolated hydrological process, primarily influenced by well-defined meteorological and topographic factors. For instance, it is generally well recognized that more incoming solar radiation, higher wind speeds, and less atmospheric moisture increase snow sublimation (Strasser et al., 2008). Thus, a model including only snow-related processes offers a sound foundation for testing the upscaling methodology.

3.2 Machine-learning Upscaling Methodology Framework

The upscaling methodology proposed and tested here has three primary steps, as shown in Figure 3.1: (1) a fine-resolution hydrological model is used to simulate detailed state and fluxes that are then spatially aggregated; (2) an ML algorithm is trained with the aggregated inputs and outputs of the fine resolution model to simulate coarse-resolution mean fluxes in any time step using only low-order statistical moments of state and forcings; (3) the ML upscaling relations are then used in a coarse resolution model to simulate the spatially aggregated state and fluxes. These coarse results may then be compared to the spatially-averaged results of the fine-resolution hydrological model for validation of the methodology. The goal is that the ML model may successfully simulate spatial mean responses of the landscape outside of the training set both in time and space and may be used to replace existing algorithms designed for use at the point scale.

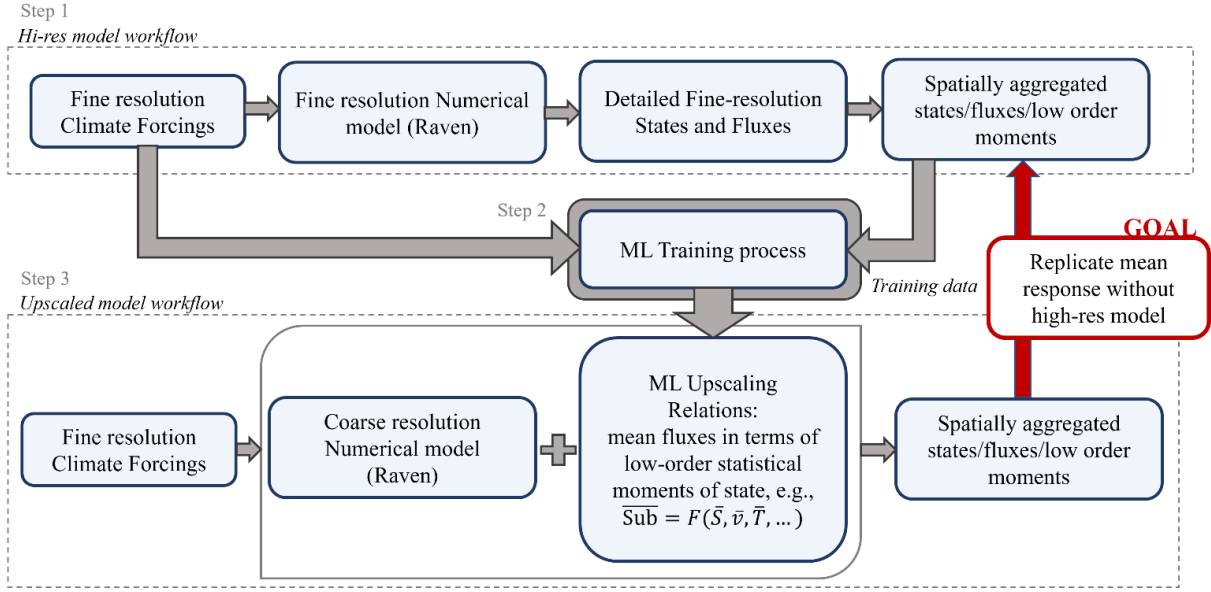


Figure 3.1: Conceptual outline of the steps used in the machine-learning based upscaling methodology.

3.2.1 The Fine Resolution Model

The development of the upscaling methodology begins with collecting an adequate quality and quantity of training data necessary for the ML model to capture spatiotemporal relationships. As in-situ measurements are unavailable at the spatial and temporal resolution required to train ML models, fine resolution model results are used. Here, the case study employs twenty-five years (1995-2020) of Daymet data (Thornton et al., 2021). Daymet is available at a 1km² spatial resolution and includes daily total precipitation, net shortwave oradiation, maximum air temperature, minimum air temperature, and vapor pressure (<http://daymet.ornl.gov/dataaccess>). Daymet also provides 1km² resolution estimates of SWE, which are used here to roughly assess the fine-resolution model. Daymet SWE estimates are generated using a temperature-index snow model, which simulates accumulation and melt based on daily temperature and precipitation inputs. Changes in SWE are computed by adding snowfall (when temperatures are below a threshold) and subtracting melt, which is estimated using a degree-day approach where melt increases with temperatures above freezing. Given the simplicity of its modeling approach, Daymet SWE estimates are expected to be approximate and likely less accurate than even the coarse-resolution Raven model outputs. As such, they are used primarily as a verification rather than for rigorous validation.

However, wind speed is not available in the Daymet database. Instead, KrigR (Kusch and Davy, 2022), a tool for downloading and statistically downscaling climate reanalysis data, is used to produce wind speed at a 1km spatial resolution and daily time step, based upon ERA5-Land climate reanalysis data. KrigR windspeed and Daymet temperature, precipitation, and relative humidity are further

downscaled to the HRU scale within the Raven hydrological model (Craig et al., 2020) using a simple elevation lapse approach, as described below.

A simple linear elevation lapse approach is applied to correct precipitation and temperature estimates at each HRU based on the specified elevation of the HRU relative to the mean forcing grid cell elevation. For example, the simple elevation lapse method for temperature corrects the HRU through the application of a lapse rate correction to the associated grid cell temperature, which can be simplified to:

$$T = T_g - \alpha(z - z_g), \quad (3.1)$$

where T is the estimate HRU temperature after correction, T_g is the temperature grid cell mean, α is the specified adiabatic lapse rate, z and z_g are the elevation of the HRU and the mean elevation of the Daymet grid cell. The adiabatic lapse rate is a global parameter within Raven and, in this case, is set to a constant value adopted from the Chernos et al. (2022) Oldman River model, which is in Alberta Canada.

For the initial test, the 955 km² Kananaskis River Basin in Alberta, Canada, was selected as the study area. The basin is characterized by its mountainous terrain, with elevations ranging from approximately 1,270 meters in the valley floors to over 3,200 meters. Its geology is predominantly composed of sedimentary rock formations, and the landscape has been significantly shaped by glacial activity (Thapa et al., 2017). For this study, landcover was simplified into three main classifications: forested areas, open areas, and lakes, focusing on the primary land features that influence snow accumulation and melt processes. The Kananaskis Valley experiences a highly variable cool, humid continental climate, marked by long, cold winters with intermittent Chinook thaws, and cool, wet summers occasionally interrupted by summer snowfalls (Alberta Environment and Sustainable Resource Development, 2014; Beckstead and Veldman, 1985).

The Kananaskis River Basin was delineated using the BasinMaker software tool (Han et al., 2023) (Figure 3.2). BasinMaker used the water survey of Canada hydrometric flow gauge (05BF025) to extract the drainage basin extent from the North American Lake-River Routing Product (NALRRP) (Han et al., 2021). From there, BasinMaker uses a DEM layer to identify key topographical features that influence hydrological flow, such as elevation, slope, and aspect. The software then generates a routing network based on the flow direction derived from the DEM, which tracks how water moves across the landscape. By analyzing these flow paths, BasinMaker can determine the boundaries of the subbasins, ensuring that each subbasin drains to a common outlet. Elevation, aspect, and landcover variables are used to generate an initial coarse discretization of HRUs within BasinMaker. Subsequently, a grid is overlain to delineate the coarse-HRUs derived from BasinMaker to a finer, near uniform resolution of 20,000 HRUs with a mean HRU area of ~0.048 km² (see inset of Figure 3.2). The HRUs represent areas where the physical characteristics and parameters are considered uniform. Mean elevation, slope, and aspect of each HRU are calculated from a 30m resolution digital elevation model. The number of HRUs is selected because, in general, the finer the resolution, the more landscape heterogeneity can be captured. Because the focus is

on snow ablation processes, the most critical source of heterogeneity is derived from topography, with some additional influence of landcover.

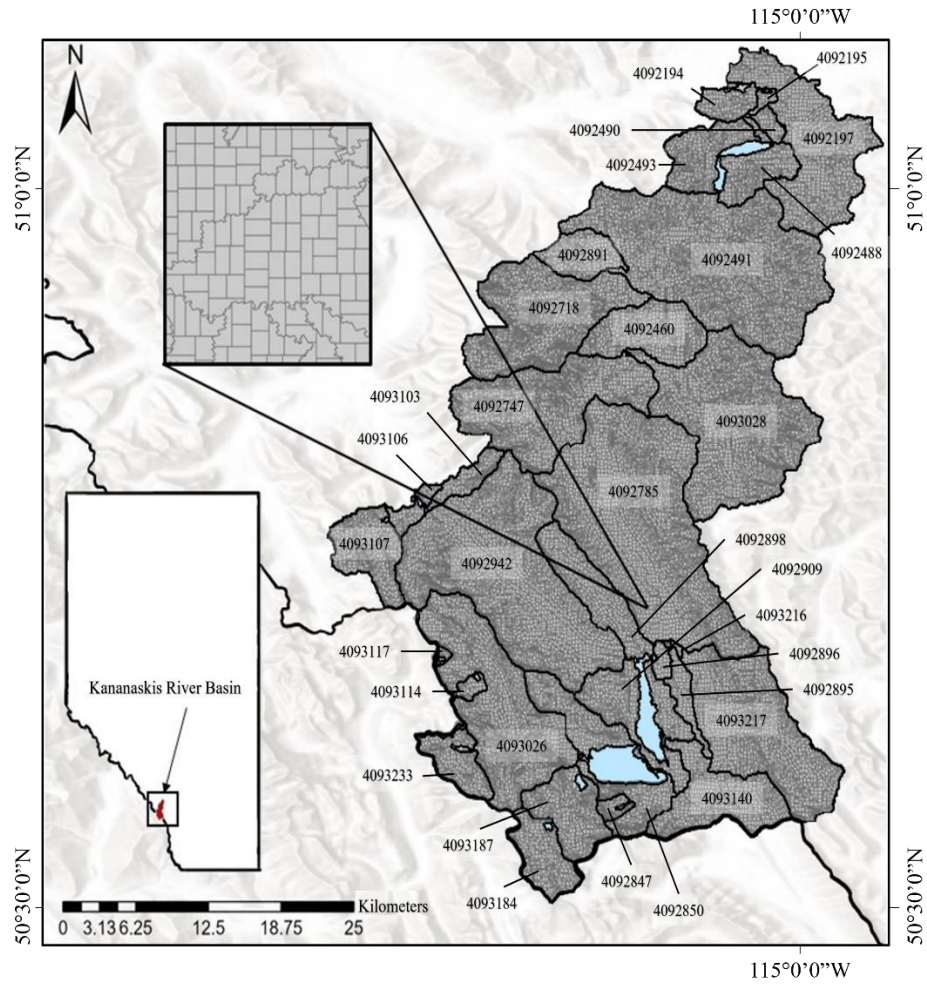


Figure 3.2: The final discretization of the Kananaskis River Basin, defined upstream of the 05BF025 hydrometric flow gauge (51°02'30" N, 115°01'55" W), resulted in 20,000 HRUs. The larger blue areas represent lakes, while the numbers correspond to the subbasin identification numbers used in BasinMaker.

A fine-resolution hydrological model built within the Raven hydrological modelling framework simulates state and fluxes for training the ML model. Raven (Craig et al., 2020) is an open-source, object-oriented software framework that supports the assessment of a wide range of model choices. Raven is chosen as it is computationally efficient and can use multiple algorithms to support the simulation of only the snow component of the hydrological cycle. For this research, Raven is configured using a two-layer snow balance model (Wigmosta et al., 1994) to simulate snow ablation via melt and sublimation, the compaction of snow, ripening and re-freezing of the snowpack, and accumulation of fresh snowfall. The rain-on-snow phenomenon is not simulated. The absence of rain-on-snow representation in the Raven

snow model may lead to underestimation of early or rapid melt events, particularly in transitional periods where warm rainfall can trigger accelerated snowmelt, an effect that may contribute to timing discrepancies in simulated ablation and runoff. The energy available for snow melt is estimated using an energy balance method.

In each HRU, the two-layer snowmelt model is used to simulate the accumulation of cold content, changes in snow surface temperature, and the evolution of both liquid and solid snow stores. The incoming melt energy initially brings the surface snowpack temperature to freezing, while the remainder is used to melt the frozen snow, which is then delegated to liquid snow until the pack is ripe, at which point it drains into the ponded water storage (Craig et al., 2020). Ripeness of the snowpack is controlled by the global snow parameter (SWI):

$$\phi_{max}^{sl} = SWI \cdot SWE, \quad (3.2)$$

where ϕ_{max}^{sl} [mm] is the maximum liquid water storage of the snowpack, and SWE is the snow water equivalent of the snowpack [mm]. The bottom (second) layer is only used when SWE exceeds the maximum SWE in the surface snow layer [mm]. The rate of sublimation [mm/d] from the snow surface, when snow is present, is calculated using the following empirical relationship (Kuzmin, 1972; Kutchment and Gelfan, 1996):

$$Sub = (0.18 + 0.098 * v_{ave}) * (P_{sat}(T) - P_{vap}), \quad (3.3)$$

where, v_{ave} [m/s] is the wind velocity at 10m, P_{sat} and P_{sat} [mb] are the saturated vapour pressure and vapour pressure, respectively, where the vapour pressure is calculated from the relative humidity and the saturated vapour pressure is estimated from average daily air temperature. When sublimation is negative, it behaves as deposition, enabling water vapour to solidify on the snowpack. The incoming energy available for melting or refreezing is calculated through the energy balance method which estimates the total snow melt through an accumulation of separate melt components,

$$M_{Melt} = \min\left(\frac{1}{\lambda_f \rho_w} ((1 - \alpha_s)SW + L_n + Q_c + Q_a + Q_r), \frac{\overline{SWE}}{\Delta t}\right), \quad (3.4)$$

where M_{Melt} is the total potential melt rate [mm/day], SW is the incoming net shortwave radiation, α_s is the snow albedo, L_n [MJ/m² 2/d] is the long wave radiation, Q_c [MJ/m² 2/d] is the convective melt energy, Q_a [MJ/m² 2/d] is the condensation or advective melt energy and Q_r [MJ/m² 2/d] is the melt energy due to rainfall. The estimates for Q_c , Q_a and Q_r are obtained using the methods of Dingman (2002) with incoming net shortwave radiation on a sloping surface calculated using Allen et al. (2006). This method requires no additional parameters; all energy estimations are based on current air and surface temperatures, as well as land/vegetation roughness heights. To ensure mass conservation, the total energy-based melt estimate is limited by the available snowpack, with $\frac{\overline{SWE}}{\Delta t}$ representing the maximum possible melt rate based on the snow water equivalent present at the beginning of the timestep.

The Raven-estimated SWE values and Daymet SWE observations are compared to one another to assess the nominal performance of the fine-resolution Raven model. For both datasets, the SWE values are spatially aggregated for each subbasin at daily intervals for the duration of the study. The Kling-Gupta efficiency (KGE) metric (Gupta et al., 2009) is used to compare the agreement between the Raven and Daymet maximum spatially aggregated SWE values. KGE is a performance metric that combines three key components: (1) correlation (r), which measures the linear relationship between the simulated and observed data; (2) variability bias (α), which compares the variability in the simulated and observed data, as the ratio of the standard deviations of the two; and (3) mean bias (β), the ratio of the means of the simulated and observed data. These components are integrated into the following equation to calculate KGE:

$$KGE = 1 - (r - 1)^2 + (\alpha - 1)^2 + (\beta - 1)^2 \quad (3.5)$$

KGE ranges from negative infinity to 1, with a value of 1 indicating perfect agreement between simulated and observed data. Generally, KGE values greater than 0.70 are considered to indicate good consistency, while values between 0.5 and 0.7 suggest moderate agreement, and values below 0.5 indicate poor performance. For all time series comparisons in this study, KGE is selected as the error metric as KGE is increasingly used in hydrological modelling studies (Shen et al., 2022). Although we assess the skill of the Raven snow model here to ensure that we are working with a reasonable model of the Kananaskis basin, the evaluation of upscaling methodology skill is independent of the Raven model performance in capturing field observations. It will instead be evaluated in its capacity to replicate results from the fine resolution model.

3.2.2 Training the Machine Learning Model

In step two of the upscaling methodology, an ML algorithm is trained on the spatially averaged state and forcing variables obtained from the fine-scale model simulation. The ML algorithm is then assessed in its ability to estimate the fluxes for a testing period. This assessment consists of three separate validation tests performed to assess the ML model's skill at estimating hydrologic fluxes at different scales and subject to variations in the means by which the model is trained.

Here, we use a RF algorithm, which consists of multiple DTs generated through a bootstrap sample of the training data and random variable subset selection (Breiman, 2001; Pang et al., 2017). Several hydrologic downscaling studies have applied RF algorithms due to their ability to handle large datasets with correlated conditional variables. RF algorithms are also resistant to overfitting, have the capacity to resolve multivariable and non-linear problems, and the ability to detect the interaction and importance of variables (Breiman et al., 2001; Kleinberg, 2000; Pang et al., 2017; Strobl and Zeileis, 2008). Chen et al. (2019) applied four ML algorithms to downscale the resolution of the total water storage and groundwater storage from 1° (110 km) and to 0.25° (approximately 25 km) and found the RF

downscaling model was the most successful. Alternative ML algorithms are not explicitly considered here, though alternative interpretable ML algorithms, such as the XGBoost model, show promise for this type of application. Similar to RF, XGBoost inherently estimates the importance of input variables aiding in the interpretability of ML algorithms (e.g., Ji et al., 2021; Rahman et al., 2020; Wang et al., 2022). LSTM models, while dominant in hydrological modelling (Anderson and Radić, 2022; Sabzipour et al., 2023), are not appropriate here as the ML model is not used directly for time series prediction, nor does it need to possess memory, as will become clear below.

The selection of input variables plays a vital role in the success of an RF model, and the omission of a significant variable will likely result in the RF model underperforming. Since the RF model is emulating the fine-resolution hydrological model, it is only capable of emulating what it is provided from the hydrological model and can, therefore, only capture physics that are present in the fine-resolution hydrological model. The state and forcing variables selected as inputs for this study were based on the variables used in the fine-resolution hydrological model to estimate basin-mean fluxes, including snowmelt and sublimation. The input and output variables used to inform the RF model are indicated in Figure 3.3. The input vector is consistent with the inputs of a hydrological model for a given time step of simulation: it includes the spatially-averaged initial state at the computational scale, plus the spatially-averaged forcings at that scale. Uniquely here, we also consider the viability of including sub-HRU variability in forcing and parameters. The initial state and forcings are used to estimate the water fluxes over the time step. We replicate the same process here, except in this methodology, the RF is used to calculate the scale-appropriate fluxes, and therefore plays the same role as explicit constitutive equations do in deterministic hydrological models. Each time step in the calibration period therefore represents a single discrete training dataset, where the spatially-averaged results from the fine-scale model provide an output vector used in training.

INPUTS		OUTPUTS	
Forcing	States	Fluxes	
Spatial mean precipitation rate [mm/d]	Spatial mean SWE [mm]	Spatial mean snowmelt rate [mm/d]	
Spatial mean temperature [C]	Spatial snow-covered area [0...1] *	Spatial mean sublimation rate [mm/d]	
Spatial mean wind speed [m/s]			
Spatial mean net shortwave radiation [mJ/m ² /d]			
Spatial mean net longwave radiation [mJ/m ² /d]			
Spatial mean snowfall rate [mm/d]			
Spatial mean snowfall temperature [C]			
Spatial Standard Deviation of Forcing Variables *			
		States	
		Spatial mean snow-covered area [0...1] *	

Figure 3.3: The contents of the input and output vector for the random forest model. All mean forcings are spatial averages over the specified computational scale. Input and output vector terms with an asterisk indicate that the input is optionally used in some subset of experiments.

Three experiments, progressively increasing in complexity, have been designed to evaluate the ML algorithm's ability to estimate spatially-averaged fluxes for groups of fine-scale HRUs. It is important to note that these experiments are not intended to assess the quality of the fine-scale snow model but rather aimed at assessing the capability of the upscaling methodology. These evaluations are designed to be insensitive to the quality of observational or meteorological data, focusing on the effectiveness of the upscaling methodology rather than the skill of the original fine-resolution model.

The experiments performed include (1) a single HRU temporal validation experiment (*Homogeneous*), (2) a spatial validation experiment on randomly grouped computational units (HETEROGENEOUS DISSIMILAR), and (3) a spatial validation experiment on similar computational units (HETEROGENEOUS SIMILAR). Landscape and topographic properties, like slope, aspect, and elevation, are not included as input vectors. This decision is based on the understanding that these properties are implicitly captured when the computation units are grouped during the clustering process. The *Homogeneous*: this temporal verification experiment (experiment 1) tests the ML model's ability to estimate the fluxes for a single target HRU without any upscaling, i.e., it simply tests how well the ML algorithm can replicate the results of the high-resolution model at a high resolution. It is expected that, because it has all of the same state information available to the original model, this *Homogeneous* performance may act as an upper bound on expected upscaling performance (because upscaling necessarily loses information about sub-region detail). A split sample validation testing dataset is used to evaluate the performance of the RF model, where 16 randomly selected years of the available year's state, forcing, and fluxes are used to train the model to estimate fluxes during the 4-year testing period (2015-2020). The training dataset reflects the seasonal dynamics of snow accumulation and melt periods, incorporating years with both above- and below-average snow conditions to represent a wide range of interannual variability. The hyperparameters are then manually constrained to produce a simplified RF model, which is subsequently verified using the four-year validation dataset (1996-2000). This tuning was done iteratively by adjusting the number of decision trees, maximum depth, and the minimum number of samples required at a leaf node incrementally, with the goal of determining how low each parameter could be reduced before a noticeable decline in performance metrics occurred, thereby producing a simplified and computationally efficient model. The final RF algorithm consists of the following parameterization: the number of decision trees was 6, the trees were pruned to the maximum depth of 12, and with a minimum of 5 samples required to be at a leaf node. Mean square error (MSE) is used to determine the optimal model parameterization. 4,000 of the 20,000 HRUs are evaluated in the *Homogeneous* experiment. The HRUs are chosen through random sampling to ensure each HRU had an equal chance of presence in the testing dataset. To evaluate whether the validation testing datasets are adequate, the mean and standard deviation of the testing mean SWE were compared against the 20,000 HRU mean SWE to ensure they are characteristically similar.

The *Heterogeneous Dissimilar* (experiment 2): this spatial validation experiment involves dividing the landscape into 12 regions based on longitude, i.e., without consideration of HRU similarity within each region. Figure 3.4 portrays how the Kananaskis basin is divided into 12 sections along the longitude, while also showing the number of HRUs and elevation within each section. The HETEROGENOUS DISSIMILAR experiment evaluates whether and to what degree natural groupings are necessary for the ML algorithm to capture the results of the high-resolution model when aggregated at a coarser resolution. The HRUs within each of the regions are randomly split into five sections where four regions are for training, and the remaining section is for testing. The region split is done at random to ensure uniformity and to increase variability within the training dataset. The RF algorithm used in the *Heterogeneous Dissimilar* and *Similar* experiments shares the same architecture as the RF algorithm derived in the *Homogeneous* experiment to avoid the RF algorithm from being specific to one area. Ideally, the RF algorithm should be transferable between experiments and areas, resulting in transferable upscaling relationships. The RF algorithm is trained on the mean training HRUs state, forcing, and fluxes to estimate the mean testing HRUs fluxes for each time step in the entire study period. The performance error metrics compare the RF estimated fluxes to the mean fluxes from the fine-resolution model. The RF model is then retrained with the addition of low-order statistical moments of forcing variables (because the goal is to simulate the mean fluxes and state, the low-order statistical moments of state will be unavailable). The forcing variables selected for the low-order statistical moments are based on the feature importance index generated by the RF algorithm, which assists in identifying significant variables that are critical in interpreting the relationships within the model. The addition of sub-HRU distributional information to the training set should assist in capturing information lost through aggregation within the stratified groups.

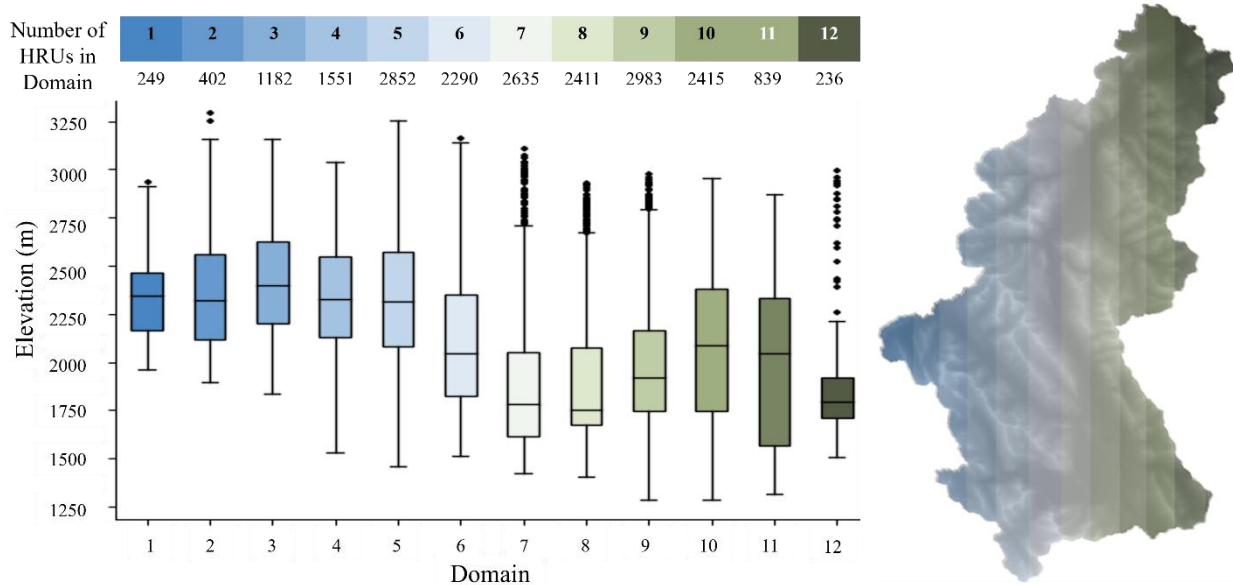


Figure 3.4: A plot of the 12 sections of the Kananaskis River Basin divided along the longitude and corresponding table indicating the number of HRUs in each section and elevation in each section.

The spatial validation experiment *Heterogeneous Similar* (experiment 3) simulates spatially average fluxes for HRU groups that share similar topographic characteristics. The HETEROGENEOUS SIMILAR experiment is intended to test the RF algorithm's ability to reproduce high-resolution fluxes for natural HRU groupings at a coarser resolution. This experiment uses a k-means cluster analysis to identify natural groupings of the HRUs based on landcover, elevation, slope, and aspect. A k-means clustering algorithm partitions the dataset into k clusters by alternatively assigning samples to the nearest centroid as measured by Euclidean distance (Likas et al., 2003). The main advantages of k-means over other unsupervised clustering methods are fast computation, low complexity, adjustable cluster membership, and the ability to process large datasets (Govender and Sivakumar, 2020). The *Heterogeneous Similar* experiment splits the HRUs in a single section into five domains of equal area. The training dataset is made up of four sections, with the remaining section in that cluster being used for testing. The ML algorithm is then trained on the 20 years of the training HRUs to estimate the fluxes for the testing HRUs for the entire study period. Due to the relatively homogenous nature of each cluster's topographic characteristics, the *Heterogeneous Similar* experiment should outperform the *Heterogeneous Dissimilar* experiment. Subsequently, low-order statistical moments of forcing variables selected by the feature importance index generated by the RF algorithm are added to the training set, and the process is repeated once more to see if the ML algorithm improves in capturing the influence of heterogeneity within the clusters.

For each experiment, a mass balance equation is deployed to replicate how the upscaling methodology would be integrated into a full model to simulate mean states using the coarse resolution

fluxes generated by the three experiments. The fluxes generated by the coarse resolution RF model are incorporated into the following simplified mass balance equation to compute SWE:

$$\overline{SWE}_i = \max(0, \overline{SWE}_{i-1} + (\bar{P}_i - \bar{S}_i - \bar{M}_i) \cdot \Delta t), \quad (3.6)$$

where $\overline{SWE}_{i-1}$ [mm] is the computed SWE value from the previous timestep, $\bar{P}_i$ [mm/d] is the mean snowfall in the timestep i within the coarse computational unit, $\bar{S}_i$ [mm/d] is the mean sublimation rate estimated by the ML algorithm, $\bar{M}_i$ [mm/d] is the mean snowmelt rate in time step i estimated by the algorithm, and Δt is the model time step, in days. The mass balance equation computes the coarse-resolution $\overline{SWE}_i$ at a daily timestep, with a non-negativity constraint applied after each step to prevent non-physical values. The resulting coarse-resolution SWE is then compared against the spatially aggregated high-resolution SWE from Raven. An agreement between the aggregated fine-resolution model and the RF model indicates that once trained, the upscaling methodology can determine the fluxes of systems at multiple scales without the need for deploying fine-resolution simulations. Additionally, a high-level of agreement between the mass balance and fine-resolution model SWE provides an alternative method of evaluation for all three experiments.

To further assess the ML upscaling methodology, the *Heterogeneous Similar* RF estimated fluxes were compared against naïvely upscaled fluxes. Naïve upscaling refers to the application of small-scale process descriptions at larger scales without accounting for scaling effects or the heterogeneity of the system (Ranjram, 2021). In this approach, the fine-scale snow model is simply applied to larger regions as if no heterogeneity were present, assuming that the same relationships apply despite the lack of consideration for spatial variability. To perform naïve upscaling, a Raven model was generated consisting of 60 HRUs that matched the extent of the domains within each cluster from the *Heterogeneous Similar* experiment. This model was used to simulate mean snowmelt, sublimation, and SWE for each HRU. The aggregated naïve upscaled $\overline{SWE}$ was compared against the mass-balance computed coarse-resolution SWE with and without distributional information and the spatially aggregated high-resolution Raven SWE for each cluster.

3.3 Results and Discussion

3.3.1 Validation

This study assesses the ability of the ML upscaling methodology in deriving constitutive relations for snow-related fluxes, enabling the simulation of large-scale hydrological responses using fine-resolution model outputs. Following the steps previously outlined, a fine-resolution Raven model

simulates a range of daily snow-related state and flux variables for 20,000 HRUs across the Kananaskis River Basin for 20 years (2000-2020). Snow-related flux and state variables reported by Raven include sublimation, snowmelt, SWE, snowfall rate, and temperature of fresh snow. SCA is determined by calculating the ratio of the area where SWE is present in the Raven model to the total area of HRUs within a subregion, however SCA is excluded from the *Homogeneous* experiment (where SCA can only be 0% or 100%). Within Raven, the lapse rate adjusts the Daymet temperature and precipitation at each HRU based on the specified elevation of the HRU relative to the local meteorological gauge.

Table 3.1: KGE values for each of the subbasins when comparing the Raven estimated daily SWE values against the Daymet SWE observations.

Subbasin ID	4092194	4092195	4092197	4092460	4092488	4092490	4092491	4092493	4092718	4092747	4092785	4092847	4092850	4092785	4092895	4092896
KGE	0.66	0.61	0.56	0.60	0.62	0.57	0.66	0.58	0.63	0.66	0.66	0.69	0.68	0.65	0.62	0.57
Subbasin ID	4092898	4092909	4092942	4093026	4093028	4093103	4093106	4093107	4093114	4093117	4093140	4093184	4093187	4093216	4093217	4093233
KGE	0.69	0.56	0.66	0.70	0.73	0.60	0.70	0.69	0.61	0.57	0.71	0.64	0.67	0.60	0.66	0.69

Due to the lack of in-situ data, the Raven estimated SWE values are compared to the Daymet SWE observations to assess the Raven model's basic performance. The SWE values for both datasets are spatially aggregated for each of the 32 subbasins at a daily interval for the duration of the study period, achieving an average KGE value of 0.64 for the 20-year period. The KGE values for each of the subbasins are presented in Table 3.1. Operating under the premise that a KGE value exceeding the 0.60 threshold denotes a moderate level of agreement, the performance metrics between the Raven model and Daymet SWE values demonstrate a sufficient level of agreement for majority of the subbasins. This threshold is commonly used in hydrological modelling studies to indicate acceptable model performance, where values above 0.7 are often interpreted as good and values above 0.90 as excellent agreement, depending on the application context and data quality (e.g., Knoben et al., 2019). In practical terms, moderate agreement suggests the model is capturing the general patterns and seasonal dynamics of SWE sufficiently, though with some limitations. To illustrate the range of performance across the study area, Figure 3.5 presents the distribution of SWE values for the fine-resolution Raven and Daymet datasets for six representative subbasins. These subbasins were carefully selected to capture the variation in model performance observed across all 32 subbasins, providing a balanced depiction of both well-performing and more challenging subbasins. However, as shown by the KGE values in Table 3.1, the results were fairly consistent among the subbasins. In the lower performing subbasins, the Raven model tends to underestimate daily SWE values compared to the Daymet dataset. The following observation is portrayed in Figure 3.5, where subbasin 4092895 exhibits a distinct concentration of SWE values closer to zero compared to the other more proficient subbasins. Although the overlaid probability density functions are intended to aid interpretation, they tend to smooth over sharp peaks and do not fully reflect the highly skewed shape of the underlying histograms. A lognormal or gamma distribution may better represent

SWE distributions and should be considered in future evaluations. However, the performance of the fine-resolution model does not hinder the upscaling model as the small-scale physics are inherently built into the upscaling methodology by the fine-resolution model, regardless of the quality of the input. Thus, the choice of the fine-resolution model is far more crucial in the upscaling methodology.

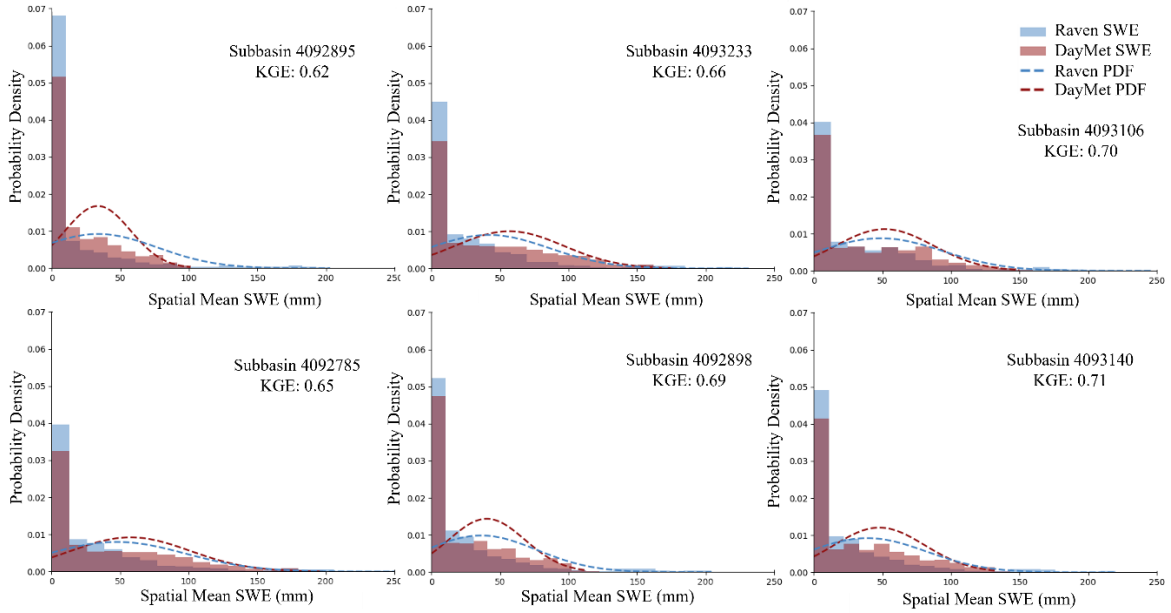


Figure 3.5: Histograms compare the distribution of the SWE values for the fine-resolution Raven and Daymet datasets. Additionally, probability density function curves of the non-zero values are overlaid to further assist in visualizing fit of the datasets.

3.3.2 Single HRU (Homogeneous)

In the *Homogeneous* experiment (experiment 1), the RF model is used to estimate the fluxes of the fine-resolution model for a set of randomly selected target HRUs in order to assess the method's skill for homogeneous systems. The *Homogeneous* experiment is applied to 4,000 HRUs, 20% of the 20,000 fine-resolution HRUs. To determine whether the testing dataset is adequate, the mean and standard deviation of the testing dataset SWE time series are compared against the SWE time series for all 20,000 HRUs to ensure they share similar variability. The 4,000 HRUs have a mean SWE of 38.7 mm and a standard deviation of 52.5 mm which is comparable to the 20,000 HRU performance metrics (mean: 38.2 mm, standard deviation: 52.0 mm). The similar variability of SWE within the datasets suggests the 4,000 HRUs are an adequate testing dataset for the 20,000 HRUs.

For the *Homogeneous* experiment, approximately 98% of the RF models for the 4,000 HRUs achieve KGE values against the testing dataset greater than 0.90 for both snowmelt and sublimation flux estimation. Figure 3.6 shows the cumulative distribution functions (CDFs) of the KGE score distribution

for snowmelt and sublimation for each of the 4,000 HRUs in the *Homogeneous* experiment. The KGE values indicate a high level of agreement between RF-estimated and the Raven simulated fluxes for homogeneous domains. To further validate the performance of the RF model, the RF estimated fluxes are used within the mass balance equation to compute high-resolution SWE at a daily timestep, which is compared against the high-resolution Raven SWE; the CDFs of the KGE score distribution for the computed SWE is also shown in Figure 3.6. The instances of lower level of agreement between the mass balance computed and Raven SWE is likely owing to snowmelt rates being estimated as zero later in the freshet season. The RF model learns that melt rates are zero during the late spring and summer months because no snow is available. However, if SWE is still present, it will persist through the summer due to the lack of RF melt. In the future, the RF model should be trained using energy availability rather than actual melt time series to address the phenomenon. The remaining 90% of the HRUs achieve KGE values greater than 0.60 for the mass balance analysis. This performance gap highlights why the mass balance estimated SWE consistently performed worse than those for sublimation and snowmelt alone. The SWE values reflect the cumulative effect of both flux estimates over time, meaning that even small errors in either component can persist and lead to greater discrepancies in the final SWE estimates. The higher level of agreement between the mass balance computed and Raven SWE suggests that the RF model is successful in capturing a substantial proportion of the variability in the dataset. Overall, the RF model successfully captures the single snow-related fluxes when scale issues are disregarded, i.e., the RF model effectively emulates Raven for this single HRU experiment.

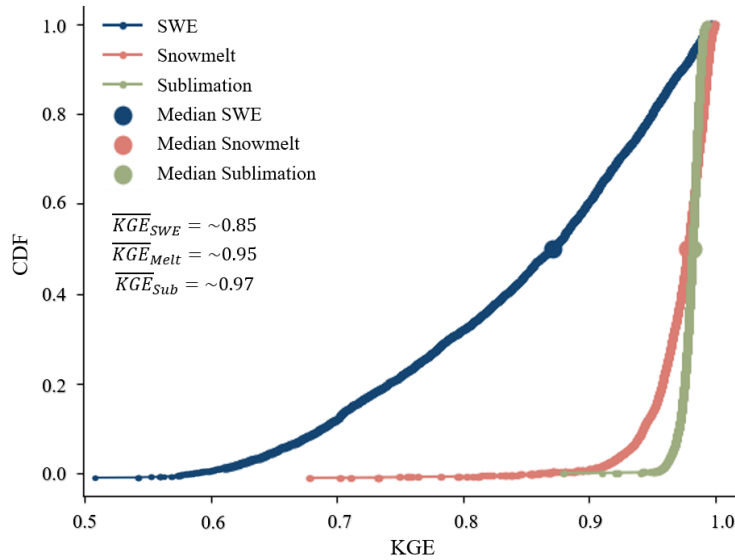


Figure 3.6: Modified KGE score CDFs of the 4,000 HRUs from the *Homogeneous* experiment for each of the three model instances.

3.3.3 Random Stratified Spatial Validation (Heterogeneous Dissimilar)

The *Heterogeneous Dissimilar* experiment (experiment 2) is designed to determine whether natural groupings are required for the ML algorithm to capture the high-resolution Raven model's results when aggregated at a coarser resolution (now heterogeneous at the computational scale of the RF). The *Heterogeneous Dissimilar* experiment does not take landcover into consideration as the region division is solely dependent on longitude. For the training and testing datasets, each of the 12 sections is divided into five sections at random. The RF algorithm is then trained on mean forcing, states, and fluxes of four of the five sections of the HRUs within the region over a 20-year period to estimate the mean fluxes for the remaining one section throughout the entire study period. Additionally, the RF model was run again excluding SCA as an input variable and instead setting it as an output for the RF model. The decision to estimate SCA as an output in the RF model, is guided by the significant impact SCA has on snowmelt. For example, when distributional information of forcing variables is later incorporated into the model, estimating SCA as an output enriches our capacity to quantify and understand the interplay between these variables and SCA.

Mean sublimation is successfully estimated in all 12 test sections, achieving a reasonably high level of agreement (KGE higher than 0.80) between the high-resolution Raven sublimation and the RF estimates in the *Heterogeneous Dissimilar* experiment. Although the RF model is successful in estimating when sublimation values fall below zero or peaks, the RF model tends to underpredict those values, as show in Figure 3.7. Similarly, the RF model demonstrated relative success in estimating SCA, as evidence by majority of the KGE values exceeding 0.75. However, the RF model SCA estimates tend to add extraneous fluctuations which, although they follow the general distribution, do not align with the Raven fine-resolution model SCA values (Figure 3.8). These fluctuations can be attributed to the inherent sparsity of the dataset and minimal variability in the SCA over a daily time step, which may have resulted in challenges in capturing the full spectrum of SCA variability.

The RF model in the *Heterogeneous Dissimilar* experiment is not as successful at estimating mean snowmelt as the homogeneous case as it is unable to capture both peak and low snowmelt values (Figure 3.7). This highlights a primary challenge in the study, which is the use of high-quality input data that is interpretable for the RF model. The datasets for snowmelt can be classified as sparse, as the variable exhibits an abundance of zeros in the dataset during months when there is no snow present. Sparse and noisy datasets can impair the performance of ML algorithms and their ability to calculate accurate estimations (Tahmasebi et al., 2020). The RF algorithm's inability to estimate snowmelt has a negative impact on the mass balance-based estimation of SWE. The median KGE for SWE is below 0.6, indicating that more than half of the sections achieved only moderate or poor agreement (Figure 3.17) Although a few sections exhibit high performance (KGE greater than 0.6), the overall spread suggests that the RF algorithm struggles to capture the system's heterogeneity when training and testing datasets are dissimilar.

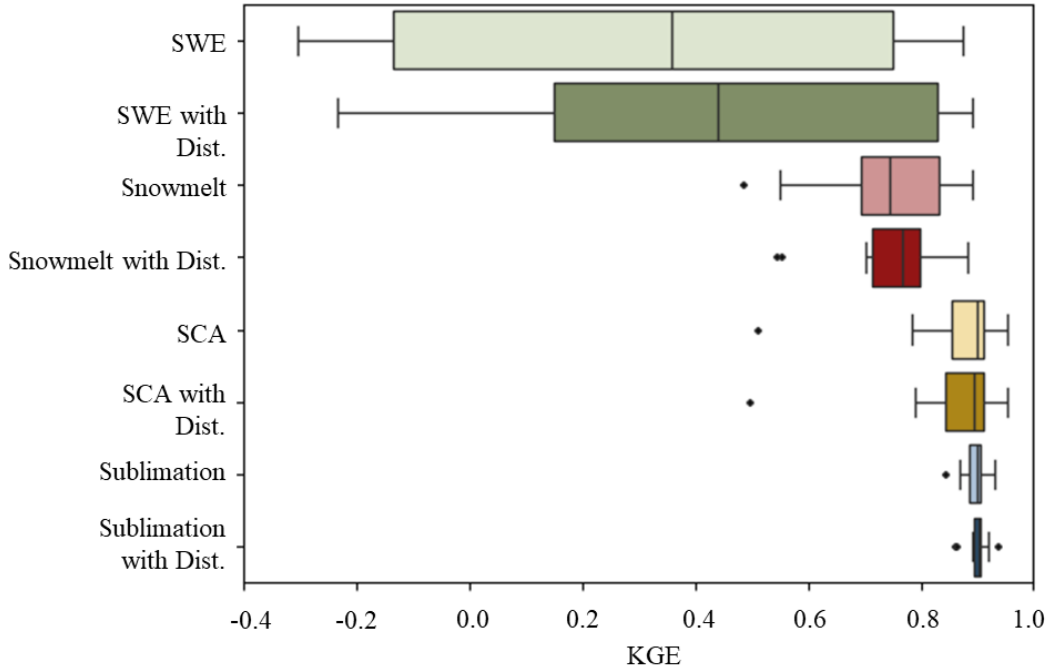


Figure 3.7: Box plot of the *Heterogeneous Dissimilar* KGE values for each section evaluating the level of agreement between the high-resolution Raven emulated and RF estimated sublimation, snowmelt, SCA, and mass balance estimated SWE.

Recognizing the impact of the system heterogeneity, the DISSIMILAR RF model is then retrained with the addition of low-order statistical moments of select forcing variables into the training dataset. The structure of the different decision trees that form the RF model is explored by permuting through the covariates to determine which forcing variables distributional information should be added to the training dataset. This provides a quantitative way to investigate the significance of covariates in the RF model, which otherwise functions as a black box. The feature importance index extracted from the RF models from the *Heterogeneous Dissimilar* experiment indicates that the temperature, net longwave radiation, and snow temperature contributes the most to reducing the RF model’s predictive error. Conceptually, these variables align well with our understanding of snowmelt and sublimation. For example, at higher latitudes, net longwave radiation can provide similar amounts of energy to snow than net shortwave radiation due to the low solar elevation (due to the cosine effect and increased scattering due to long atmospheric path lengths). This effect is magnified in mountains due to shading and net longwave emissions from the complex topography (Sicart and Pomeroy, 2006).

The addition of distributional information to the training dataset improves 9 of the 12 sections KGE values for SWE, with 5 sections exhibiting SWE KGE values greater than 0.6 (Figure 3.8). Figure 3.8 depicts a cross-section of section 6’s daily spatially aggregated high-resolution Raven snowmelt, sublimation, SCA, and SWE against the computed coarse-resolution RF outputs with and without distributional information. Section 6 is chosen due to its KGE values aligning closely with the median

KGE values calculated across all sections. This selection ensures a representative and consistent performance benchmark within the sections. The addition of variable distributional information improves the estimation of SWE as it assists in capturing reducing the over or underestimation of SWE.

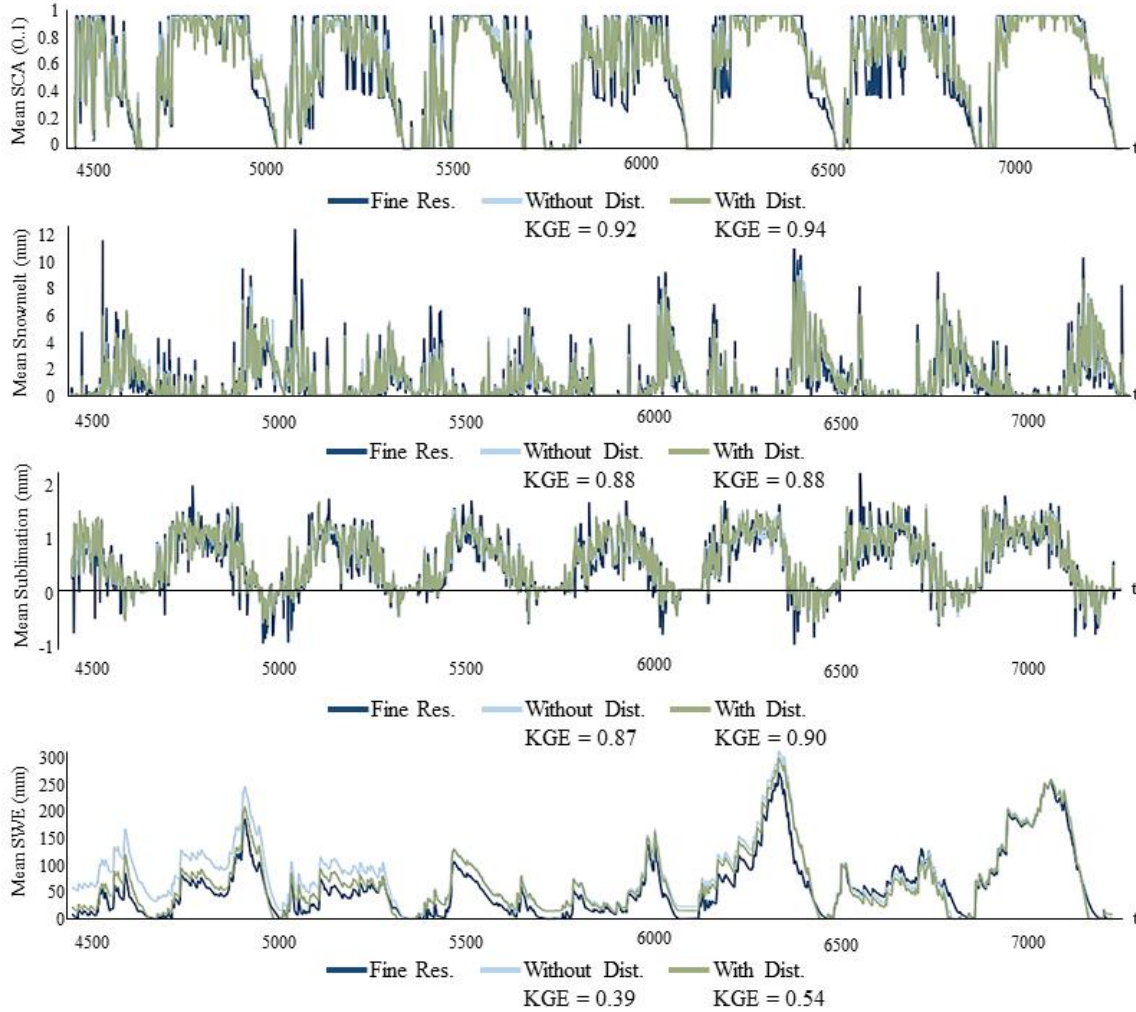


Figure 3.8: Mean SCA, snowmelt, sublimation, and SWE in section 6 comparing the outputs of the Raven model and mass balance with and without inclusion of distributional information over 3000 daily timesteps.

The *Heterogeneous Dissimilar* experiment represents the most challenging scenario, where the upscaling methodology is tested on a dissimilar section from which it was trained on—that is, it is tested on unseen locations. The addition of the distributional information provides the foundation that the RF model may be more capable of capturing system heterogeneity if some lower order statistics of forcings are accessible. However, there are still regions with minimal agreement between the spatially aggregated high-resolution Raven and coarse-resolution SWE with and without distributional information. Analysis of elevation box plots in the *Heterogeneous Dissimilar* experiment (Figure 3.4) reveals a prevalent

correlation that sections with higher agreement levels tend to showcase less variability in elevation data and fewer outliers, while those with more outliers and increased variability often correspond to lower KGE values. It is noteworthy that, despite elevation data not being directly inputted into the RF model, its implicit influence is captured through subbasin subdivision along the longitude. This indirect integration accents the significance of considering elevation dynamics. Therefore, the outcome of the *Heterogeneous Dissimilar* experiment results indicates the necessity for some level of similarity between the training and testing datasets.

In the initial *Heterogeneous Dissimilar* experiment, the RF model was deliberately restricted to a simplified architecture of only 6 decision trees. This was done to support the future integration of the ML-derived upscaling relationships into existing hydrologic models, where interpretability and computational efficiency are important. While the *Heterogeneous Dissimilar* experiment may have struggled to capture the system's heterogeneity due to dissimilar training and testing datasets which arise from variability in topography and land cover in the domains, another potential limitation is the simplified RF architecture. With only 6 trees, the RF model may not have been able to fully capture patterns in the sparse and noisy snowmelt data, which contains many zero values during snow-free periods. To assess whether the limited architecture impacted performance, the RF model was rerun using 100 trees and a maximum depth of 20. As shown in Figure 3.9, both versions of the model yielded similar KGE distributions and median values, suggesting that increasing the number of trees offered only minor improvements. The deeper model slightly reduced noise during transition periods (e.g., accumulation or melt onset), but overall, the simplified RF still captured the key relationships needed for upscaling. This comparable performance is likely due to the dominant patterns captured in the training data for the snow-related fluxes which are primarily governed by dominant topographic controls and season patterns, which the simplified RF model could effectively represent. These findings support the use of the simplified RF model in future applications, where maintaining a balance between model interpretability, computational efficiency, and predictive accuracy are vital.

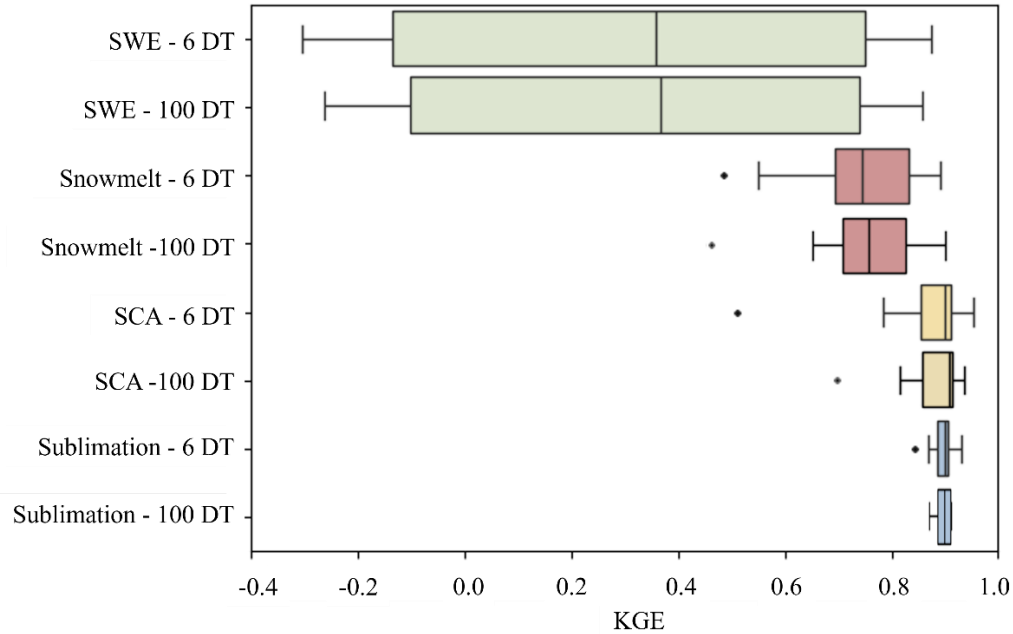


Figure 3.9: Box plot of the Heterogeneous Dissimilar KGE values for each section evaluating the level of agreement between the high-resolution Raven emulated and RF estimated sublimation, snowmelt, SCA, and mass balance estimated SWE. Comparisons are shown for RF models using either 6 or 100 trees.

3.3.4 Similar HRU (Heterogeneous Similar)

The *Heterogeneous Similar* experiment (experiment 3) evaluates the RF algorithm capacity to estimate spatially average snow-related fluxes when training and testing sets share similar characteristics. To identify similar HRUs, the landcover, elevation, aspect, and slope data for all the HRUs were first classified into 12 spatially discontinuous clusters using the k-means algorithm. The optimal number of clusters was derived through an elbow method analysis, which identifies the point where adding more clusters yields diminishing returns in explained variance. Figure 3.10 shows the spatial distribution of the clusters, number of HRUs, variability of elevation, and landcover percentages within each cluster. Each cluster is divided into five domains, four of the domains are used as a training dataset and the remaining domain as a testing dataset. The RF algorithm is then trained on the training domains daily mean fluxes over a 20-year period to estimate the daily mean fluxes for the one testing domain over the entire study period.

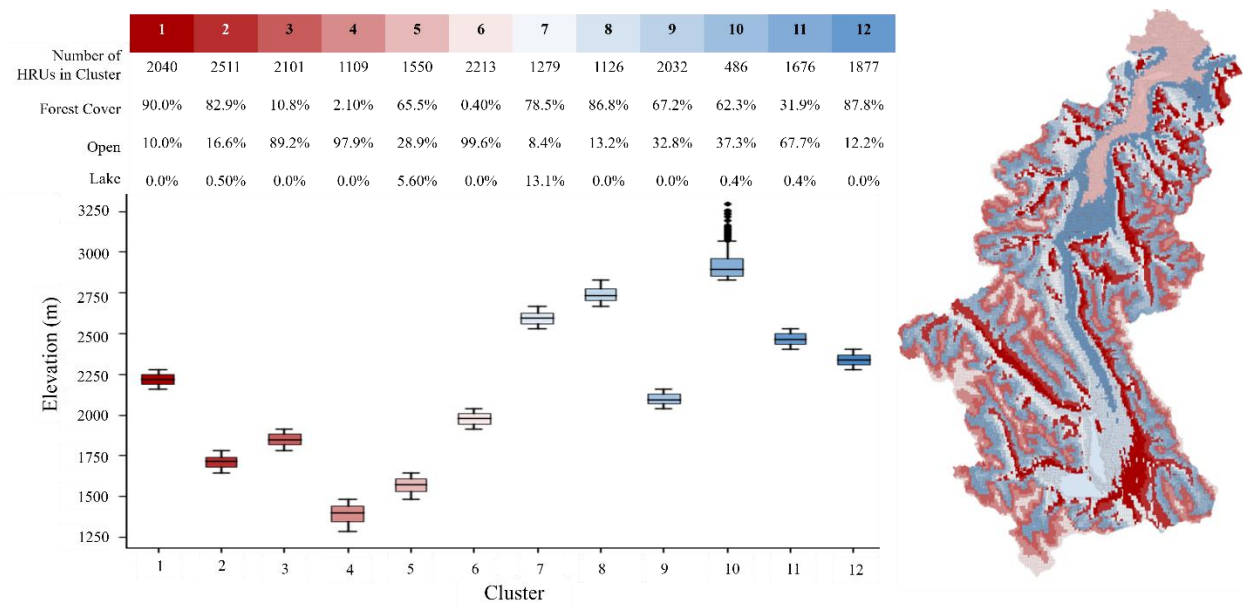


Figure 3.10: A plot of the 12 k-means clusters of the Kananaskis River Basin and a corresponding table indicating the number of HRUs in each domain, elevation, and landcover percentages in each cluster.

The RF model's predictive performance for the *Heterogeneous Similar* experiment is portrayed in Figure 3.11. Similar to the findings in the *Heterogeneous Dissimilar* experiment, each cluster exhibited a KGE value surpassing 0.80 for sublimation. However, it is noteworthy that the RF model continues to underestimate the peak sublimation values. The SCA estimates improved, particularly due to the reduced noise in the RF models estimates. Snowmelt performance notably improved in the *Heterogeneous Similar* experiment, with all 12 clusters showcasing a median KGE value of 0.93 in the presence of distributional forcing variables. In contrast, the *Heterogeneous Dissimilar* experiment yielded a significantly lower median KGE of 0.74. Particularly, the RF algorithm demonstrates a discernible enhancement in the estimation of zero snowmelt values and peaks. Furthermore, 9 out of the 12 clusters successfully computed coarse-resolution SWE on a daily-basis (KGE greater than 0.6), without the need for fine-resolution data, as shown in Figure 3.10. While there have not been notable improvements in the estimation of zero snowmelt values for some clusters, this improvement is not uniformly evident across all the clusters. The lower performing clusters are likely owing the RF algorithms failing to capture the persistence of summer snow, resulting in the RF model producing noise in its estimates.

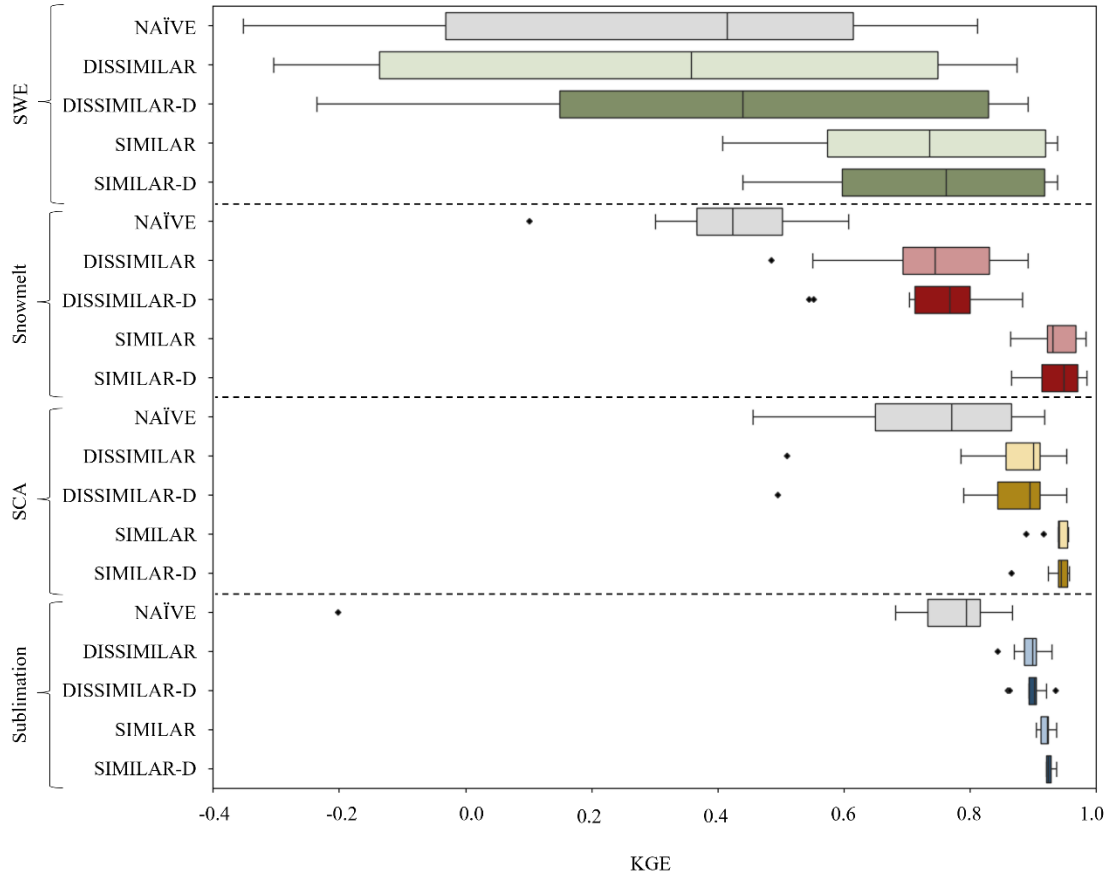


Figure 3.11: Illustrates the KGE values for all clusters, evaluating the agreement across various variables, including high-resolution Raven emulated SWE, Raven naïvely upscaled SWE, and RF-estimated snowmelt, sublimation, SCA, and mass balance estimated SWE in both *Heterogeneous Similar* and *Dissimilar* scenarios with and without distributional information.

The characteristic sparsity of data, particularly evident in variables like snowmelt and sublimation rates during non-winter months, presents a notable challenge for ML algorithms in this study. The abundance of zeros significantly impacts the model's learning process, skewing its understanding and prediction capabilities towards non-events rather than capturing the nuanced dynamics of hydrological processes. The prevalence of zeros can lead to an overemphasis on null values, potentially diminishing the model's ability to accurately predict the occurrence and magnitude of snow-related fluxes when they do happen, as seen in RF models consistent inability to estimate peak sublimation values. This challenge underscores the need for quality input data which may be accomplished through tracking energy availability rather than actual melt. Furthermore, the selection of an alternative loss functions may assist in accounting for the data sparsity. This study uses a simple, uniform RF model structure across all experiments to ensure the model is not overly tailored to the specific characteristics of the Kananaskis River Basin. This simplification supports our aim for the upscaling relationships to be transferable across

basins that share similar characteristics, which is beyond the immediate scope of this study. In the future, a critical evaluation of the RF model's architecture, particularly the level of complexity of the RF model while still maintaining a level of transferability will be beneficial in addressing the challenges posed by data sparsity.

The *Heterogeneous Similar* experiment yields consistently greater KGE values in comparison to the *Heterogeneous Dissimilar* experiment. This outcome further reinforces earlier observations, affirming that the incorporation of natural groupings aids the RF algorithm in effectively capturing the high-resolution Raven model's outcomes when aggregated at a coarser resolution. However, the RF model's performance, when trained and tested exclusively with spatially averaged values in the *Heterogeneous Similar* experiment, is (unsurprisingly) not as successful as the *Homogeneous* experiment. This is apparent in the lower median KGE values for sublimation, snowmelt, and mass balance SWE estimates across all clusters (0.92, 0.93, and 0.74) compared to the higher median KGE values observed in the *Homogeneous* experiment (0.95, 0.97, 0.85). This is reflective of the information loss associated with “hiding” distributional data from the RF model – because it doesn’t have access to (e.g.) the variance of net shortwave radiation, it can less capably represent the aggregated response. Recognizing the impact of this information loss, the RF model is then retrained with the addition of standard deviation of select forcing variables into the model input vector. The analysis of covariate importance in the RF models derived from the *Heterogeneous Similar* experiment reveals that SCA and SWE play a predominant role in minimizing the predictive error of the RF models across most clusters. However, it is important to note that this study concentrates on the incorporation of distributional information related to the forcing variables, thus, temperature and net longwave radiation are selected instead.

Along with the addition of distributional information, the naïve upscaling model uses the testing HRUs present within each of the 12 clusters and aggregates the SWE values through a coarse resolution Raven model. The naïve Raven model tends to perform poorly with a median KGE of 0.49, which is notably lower than the *Heterogeneous Similar* KGE’s both with and without distributional information (0.74, 0.76). The naïve Raven model frequently under or overestimates mean SWE values, as it overlooks the potential influence of scaling effects and the inherent heterogeneity within the system. This omission leads to both overestimation and underestimation of SWE, as the model fails to account for the effects of varying topographic and landcover characteristics.

However, the naïve model often outperforms the RF model estimates when SWE drops to zero, typically occurring in later spring and warmer months. The RF model's inability to accurately estimate these sharp drops to zero stems from an inherent limitation of the RF algorithm. Similar to many ML algorithms, the RF model operates on the principle of fitting smooth and continuous functions to the input data. Consequently, it may struggle to precisely represent abrupt, discontinuous changes, such as sudden drops to zero. As previously discussed, a potential solution for addressing this phenomenon in

future iterations could involve training the RF model using energy availability data rather than actual melt time series.

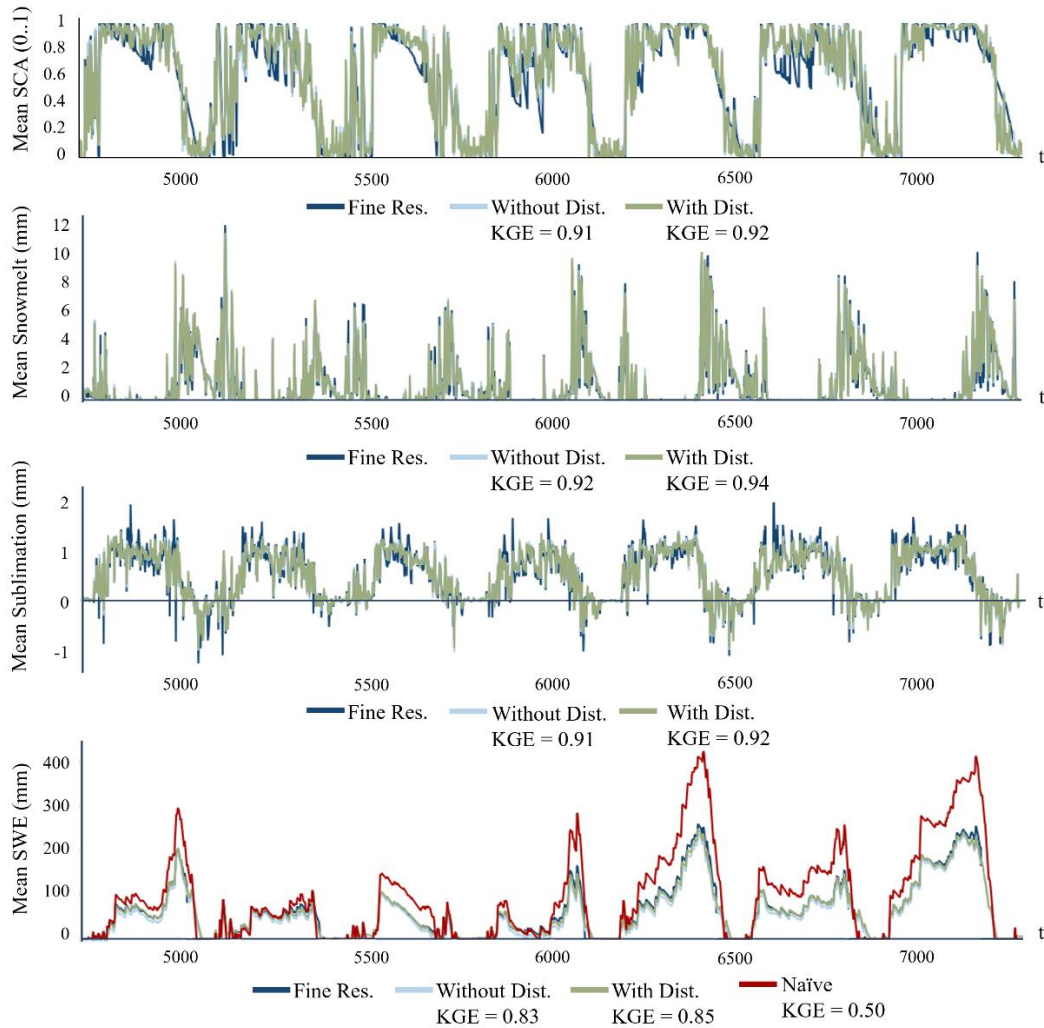


Figure 3.12: A plot of 3000 daily timesteps of cluster 8 outputs for mean SWE comparing the outputs of the Raven fine-resolution model, naïve upscaled model, and mass balance without and with distributional information.

With the addition of the distributional information to the training dataset, 5 of the 12 clusters KGE values for the mass balance equations improved by more than 0.03 and 7 remained the same. Figure 3.12 illustrates 2500 daily timesteps for cluster 8's outputs for the five SWE estimation methodologies. Cluster 8 was selected as its median KGE values aligns with the median KGE *Heterogeneous Similar* for all clusters, with and without distributional data. The similarity in medians suggests that cluster 8 serves as a representative case for all clusters as the median is a robust measure of central tendency. The addition of

distributional information of forcing variables into the RF algorithm has proven beneficial by capturing previously underestimated snowmelt and sublimation values. However, the minimal change in performance for some clusters may be attributed to the study focus on incorporating only distributional information of forcing data. As previously mentioned, SWE and SCA emerged as the most influential variables during covariate analysis across many clusters. Nonetheless, snowfall temperature, precipitation, and windspeed also demonstrated significance in certain cases. This highlights the complex interplay of factors influencing snowpack dynamics. Future research efforts should broaden their scope to include the distributional information of state variables alongside forcing variables. Additionally, customizing the selection of distributional variables for each cluster could prove instrumental in effectively capturing the inherent heterogeneity of the system. This tailored approach is likely to yield more nuanced insights and further assist in capturing the heterogeneity of the system. Furthermore, the computational cost required to run a Raven model using the ML upscaling methodology for 12 HRUs is significantly lower than those required for the fine-scale model with 20,000 HRUs over 20 years. This efficiency advantage demonstrates the potential of the ML upscaling methodology to deliver substantial computational savings in large-scale hydrologic modelling.

3.4 Conclusion

This study set out to develop a ML upscaling methodology to generate ML-estimated constitutive relations between systemwide storage and fluxes to simulate coarse-resolution mean hydrologic system response without the need for fine-scale simulations. In the *Homogeneous* experiment, the RF model accurately captures snow-related fluxes when scale issues are not taken into consideration. As the experiments progressively increase in complexity, the *Heterogeneous Dissimilar* experiment tests the upscaling methodology on a dissimilar domain from its training dataset. The addition of distributional information helps the RF model capture system heterogeneity, but there are still areas of disagreement between the high-resolution and coarse-resolution SWE. This experiment underscores the importance of having some level of similarity between the training and testing datasets, which can be achieved through natural groupings, as shown in the *Heterogeneous Similar* experiment. Given the relatively uniform topographic and landscape features within each cluster, the RF model performs better when regions are defined by similar characteristics, as seen from comparison between *Heterogeneous Similar* and *Heterogeneous Dissimilar* experiment. Consequently, the inclusion of low-order statistics of forcing variables identified by the RF algorithms features importance index assists in capturing the system heterogeneity for 5 of the 12 clusters. The limited improvement for the remaining clusters can be attributed to the selected forcing variables not being notably influential in those clusters. Future research should consider incorporating distributional information for state variables and tailor the selection of distributional variables for each cluster to better capture the system's heterogeneity. It is clear that these

techniques are much more skilled at capturing fluxes than derived states, in part because snow state has a threshold at zero.

The ML upscaling methodology, with and without the inclusion of low-order statistics of forcing variables, represents a new approach for upscaling small-scale snow-related processes when compared to traditional upscaling strategies such as effective parameterization the naïve upscaling model, especially in the case where there are few tuneable parameters. The ML upscaling methodology's superior performance compared to the naïve upscaling model emphasizes the significance of capturing system heterogeneity during upscaling and exposes an issue with negligent scale effects and sub grid heterogeneity in snow models. Furthermore, the suggested methodology presents a novel application of ML algorithms beyond their conventional use in hydrology. Rather than identifying and mimicking patterns between time series of forcing inputs and predicted outputs to improve forecasts, the ML algorithm infers scale-dependent constitutive relations from the fine resolution models. It is anticipated that the ML upscaling methodology will prove useful for understanding the role of variables at various scales, for improving the representation of system heterogeneity in hydrologic models, and for improving the computational efficiency of models.

Chapter 4

Assessing machine learning-identified upscaled constitutive relationships for snow accumulation and ablation in mountainous terrain

4.1 Introduction

Scale is a fundamental challenge in hydrology, stemming from the primary constraint of only being able to observe hydrologic processes over a limited range of spatial extents. Conventionally, observations in the lab and field are distilled into equations defining relationships which describe the observed phenomena at the (typically very local) observational scale. The challenge then becomes how we can transfer these small-scale relationships to the watershed scale in a computationally efficient and sufficiently accurate way. This is the very core of computational upscaling. Unfortunately, the process of upscaling is not straightforward: it requires an understanding of small-scale heterogeneity and/or defensible simplifications of this heterogeneity. For example, snow dynamics play a vital role in hydrological regimes, with modelling of snow accumulation and melting processes contributing to the estimation of seasonal snow depletion within a watershed. This, in turn, aids in the prediction of runoff patterns (Barnhart et al., 2016; Viviroli & Weingartner, 2004). However, the spatial distribution of snow, particularly in mountainous areas, exhibits spatial heterogeneity and variability at multiple scales, primarily driven by the complex interactions between atmosphere, vegetation, and terrain (Sturm & Wagner, 2010; Trujillo, Ramírez, & Elder, 2007).

The dynamics and energetics of snow vary over multiple scales, ranging from point scale to catchment-scale patterns of accumulation and ablation. The variability at each scale is driven by topography, vegetation, wind, and incoming energy that govern snow accumulation and fluxes. As the interactions among these drivers change across different scales, the variability of snow accumulation and ablation shifts accordingly. The spatial heterogeneity of snow depth, density, and albedo is primarily driven by spatially variable topographic, landcover, and meteorological characteristics (Revuelto et al., 2020). For instance, meteorological characteristics play a pivotal role, as forcing variables, such as temperature, radiation, and precipitation directly affect the amount of local snow accumulation and ablation (Thackeray et al., 2019). Topographic characteristics, including elevation, slope, and aspect, dictate the distribution of snow within a region, with higher elevations retaining snow for longer periods due to colder temperatures (Pomeroy and Gray, 1995; Saydi and Ding, 2020). Slope and aspect influence the amount of incoming solar radiation upon the snowpack, with south-facing slopes melting faster than north-facing ones, thanks to greater exposure to solar radiation (López-Moreno et al., 2014). Landcover, including the type of vegetation, modulates snowmelt and melt runoff rates. For instance, forested areas tend to experience slower snowmelt due to the shading effect of foliage which reduces the amount of incoming solar radiation. Conversely, barren or urban areas generally experience accelerated melt (Essery et al., 2008).

Various computational upscaling techniques exist to capture subgrid variability, such as the use of effective parameters (Jana and Mohanty, 2012) or stochastic approaches (Kavvas, 2003; Ohara et al., 2008, 2014). These techniques have been extensively catalogued in Ranjram (2021). One technique, effective parameterization is a commonly used method whereby parameters and processes inside each computational element are treated as if they were homogeneous and that the small-scale descriptions can be applied to the entire element. The microscale parameters are replaced by effective parameters at the macroscale, across the entire domain, where the parameterization is hoped to compensate for the lack of explicit treatment of heterogeneity. The use of effective parameters is often not optimal because nonlinear heterogeneous systems do not necessarily behave in the same way as homogeneous systems with effective parameters (Kavvas et al., 2013). Subgrid variability is represented in stochastic upscaling approaches by probability density functions (PDFs), which typically represent the distribution of subgrid parameter values within the computational element (Kavvas, 2003). For instance, Ohara et al. (2008) integrated PDFs into a distributed snow model using the Fokker-Planck Equation to capture subgrid variability in snow accumulation and melt processes. The disadvantage of such stochastic approaches includes their computational complexity for fine-scale simulations over large geographical areas and their reliance on numerous assumptions about cross-correlation of variables and parameters (Clark et al., 2017).

Chapter 3 developed and tested a ML upscaling methodology to estimate spatially averaged snow states and fluxes in heterogeneous computational units without requiring explicit knowledge of local detail. The process involves several steps. First, a fine-resolution hydrological model is used to simulate

detailed states and fluxes, which are then spatially aggregated. Next, an ML model is trained using the aggregated inputs and outputs of the fine-resolution model, incorporating low-order statistical moments of state and forcing variables as inputs and outputs to capture distributional information that is typically lost through aggregation. The ML model then simulates the coarse-resolution mean fluxes over an independent study period and/or independent model subdomain. These were then validated against the spatially averaged results of the fine-resolution hydrological model. This methodology was found to successfully represent the impact of heterogeneity within the system without simulating it explicitly and has potential for identifying key concepts on scaling relationships for snow ablation. For instance, what level of heterogeneity may be ignored? Is spatial variability in aspect more important than variability in forest cover? Embedded in the ML model are effective scale appropriate constitutive relations for snow processes. This study aims to examine how the ML upscaling methodology of Chapter 3 can be employed to identify these dominant factors, thereby enhancing our understanding of scale effects in snow hydrology.

ML models are recognized to have superior predictive performance when compared to classic statistical models, such as multiple linear regression, as ML can identify effective relationships between input and output data even if the underlying mechanism described by the data is complex, nonlinear, and unknown (Raghavendra & Deka, 2014). However, the complexity of a fully trained ML model makes it difficult to clearly interpret or quantify the relationships between variables in the data. According to Nearing et al. (2021), this concept "is more of a testament to a lack of inspection, rather than a limitation of the models themselves," implying that we must learn how to decipher the information contained in the ML models. Interpretable ML methods have recently been used to explain what the ML models are learning and to derive new insights from complex natural systems (Guidotti et al., 2019). These include methods like Gini feature importance (Louppe et al., 2013), SHAP (Lundberg & Lee, 2017), and accumulated local effects (ALE). These approaches can identify significant variables, visualize their nonlinear effects, and quantify the controls. For instance, Wang et al. (2024a) used the SWAT model to simulate watershed runoff providing critical inputs for an XGBoost model. The study then employed SHAP to identify interactions between land use and meteorological factors, such as temperature and humidity, that influence runoff dynamics across different land use periods. Most earlier studies that focused on the influence of catchment characteristics, such as land use, soil type, and topography, employed simple regression approaches (Padrón et al., 2017; Xu et al., 2013; Zhang et al., 2018), while ML methods were used primarily as black-box models for predictions (Sinha et al., 2019). However, interpretable ML techniques have begun to shift this paradigm, enabling researchers to explore how catchment characteristics interact in complex, nonlinear ways and to extract more meaningful insights from these predictive models. In hydrology, cross-validation techniques and diagnostic tools are essential for validating models and providing new insights. There is the potential to apply methods, such as sensitivity analyses, to further assist interpretable ML methods to better understand input contributions

and translate the information within the model to identify key controls on scaling (Sundararajan et al., 2017).

This work aims to gain a better understanding of the scaling relationships between snow fluxes, states, and landscape properties over a range of scales through an analysis of scale-appropriate constitutive relations implicitly derived through the ML upscaling methodology developed in Chapter 3. Subsequently, the relationships derived from one study area will be extrapolated and tested in a different geographic location exhibiting comparable topographic and landcover characteristics, assessing the relationship's transferability. The objectives of this study are to use the ML models to 1) discover the conditions under which emergent constitutive relations are (or are not) valid, 2) determine what the key controls are on these upscaling relations, and 3) validate the transferability of constitutive relations to areas that share similar terrain characteristics.

4.2 Methods

4.2.1 Study Area

This study will concentrate on two primary study areas: the Kananaskis River Basin and the Oldman River Basin. The Kananaskis River watershed (955 km²) is located in the eastern slopes of the Rocky Mountains in Alberta, Canada, along the provincial boundary with British Columbia. As a region with a typical semi-arid climate, the study area experiences long, cold, and snowy winters, along with warm, dry summers. Seasonal snow generally begins in October, forming a stable cover by mid-December. The SWE peaks in mid-February or early March, with the snow usually melting completely by the end of May. Kananaskis is characterized by high topographic variability, with elevation ranging from 1271 to 3205m, and high spatial heterogeneity, with a variety of landcover types, including grassland, forest, lakes, alpine, and glaciated areas which often vary in according to elevation. Due to the combined effects of climate conditions, topographic features, landcover types, and other factors, the snow cover distribution in this region exhibits complex spatiotemporal heterogeneity. The Kananaskis River basin discharges into the Bow River which flows through the Alberta foothills onto the Prairies to join with the Oldman River.

The Oldman River Basin, covering an area of 3,152 km², serves as the second study site in this research to assess the transferability of the constitutive relations established for the Kananaskis River Basin. Located in southern Alberta, with headwaters originating in the Rocky Mountains and approximately 270 km separating the two study areas, the Oldman River Basin was chosen for its similarities to the Kananaskis River Basin, including a semi-arid climate, comparable elevation range (1,268 to 3,216 m), and analogous landcover. Figure 4.1 illustrates the geographic locations of Kananaskis

and Oldman within Alberta, alongside elevation and landcover percentages. While both basins are predominantly forested, Kananaskis features a higher proportion of alpine terrain, whereas the Oldman basin extends further into the lowlands, includes more grassland and shrubland areas.

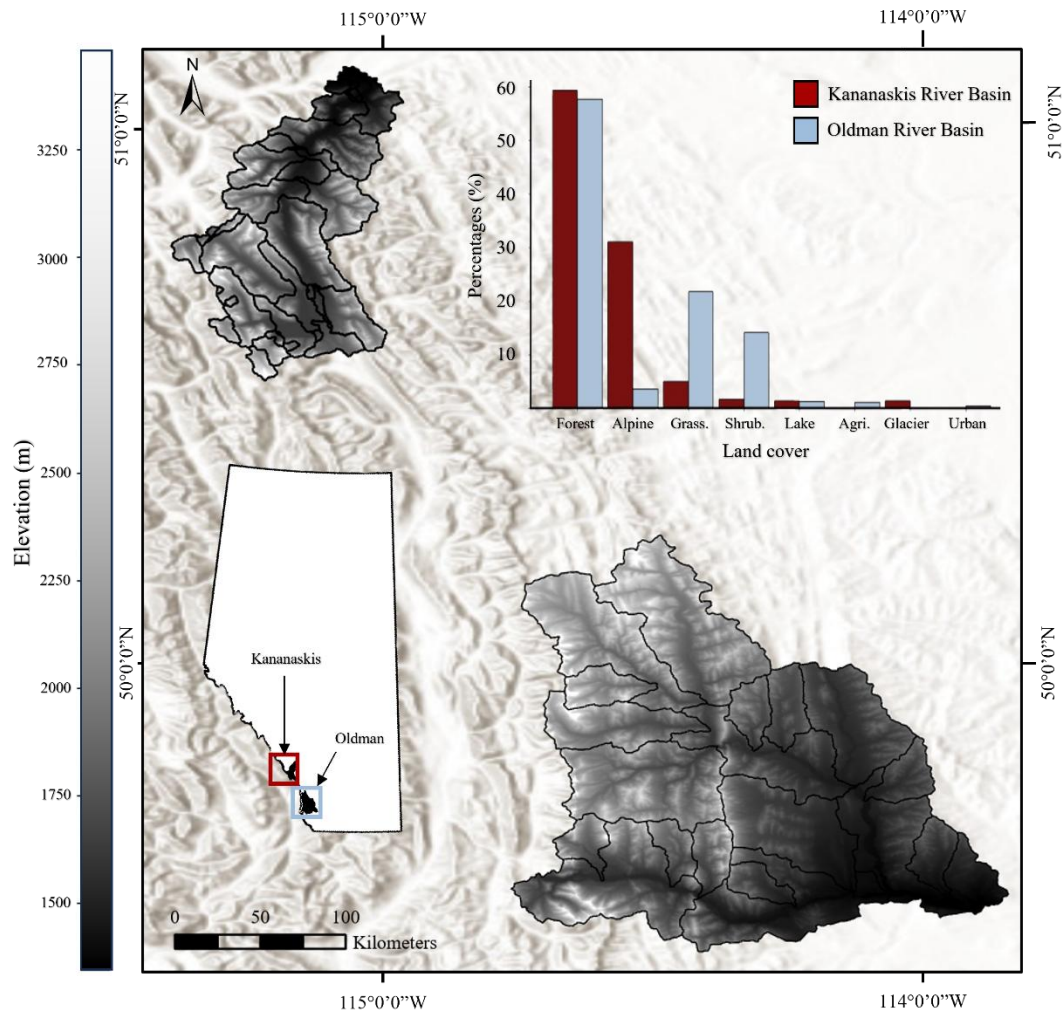


Figure 4.1: The geographic location of the Kananaskis and Oldman watersheds in Alberta, Canada, highlighting the topographic relief of the study area. The bar chart inset illustrates the percentage distribution of the landcover types within each watershed.

4.2.2 Datasets

Forcing data for both the Kananaskis and Oldman models was acquired through the Daymet reanalysis product (Thornton et al., 2021) from 1995 to 2020, available at a 1km² spatial resolution. Daymet is an interpolated dataset that generated gridded estimates of daily weather parameters, including total precipitation, shortwave radiation, maximum air temperature, minimum air temperature, vapor pressure, and SWE, by applying statistical modelling techniques to ground-based meteorological observations. Within the Raven model, a simple linear elevation lapse approach is used to adjust temperature and precipitation according to the HRU's elevation relative to the mean elevation of each

forcing grid cell at the HRU scale ($\sim 0.035 \text{ km}^2$). Wind speed at a 1 km^2 resolution was derived using Windmapper, a wind field downscaling tool developed by Marsh et al. (2023) to improve the accuracy of wind speed and direction estimates across complex terrains. Windmapper utilizes the mass-conserving WindNinja model (Wagenbrenner et al., 2016) to capture terrain-induced wind variations by precomputing a library of wind speedup and direction perturbation maps, enabling efficient downscaling of mesoscale wind fields like the ERA5 u- and v-wind components (Hersbach et al., 2020). Windmapper results are then interpolated to the HRU scale within the Raven hydrologic model through inverse distance weighting (Craig et al., 2020).

4.2.3 Fine-resolution Hydrologic Model

The ML upscaling methodology is trained on a fine-resolution snow model built within the Raven hydrological modelling framework for the Kananaskis River Basin. Raven (Craig et al., 2020) is an open-source, object-oriented software framework which can be configured to simulate only the snow component of the hydrological cycle. For this study, the Kananaskis model was developed using the Magpie Workflow (Burdett et al., 2024), which is an open-source Python/R workflow that provides structured steps for generating the necessary input files for Raven. Magpie uses the BasinMaker software tool (Han et al., 2023) to discretize the Kananaskis study area based on elevation and landcover layers, the Subdivide Polygon tool from ArcGIS (Esri, 2024) split the BasinMaker-derived HRUs into approximately 27,000 HRUs relatively equal in area ($\sim 0.035 \text{ km}^2$). The spatial discretization used here differs from Chapter 3, where a simplified landcover layer consisting of forest, open area, and lakes was intentionally applied to evaluate the ML upscaling methodology under more controlled conditions. In contrast, Chapter 4 incorporated a more detailed land cover layer to better reflect real-world landscape variability. Raven is configured using the same two-layer snow balance model based up that of Wigmosta et al. (1994) as described in Chapter 3. This configuration aims to simulate snow ablation via melt and sublimation, compaction of snow, ripening and refreezing of the snowpack, and accumulation of fresh snowfall. The energy available for snowmelt is estimated using an energy balance method. The energy balance estimates potential melt based on convective/conductive heat transfer, and the energy content of rainfall (Craig et al., 2024).

4.2.4 General Experimental Design

A series of experiments are described to evaluate the ML algorithm's ability to estimate spatially averaged fluxes for large regions under various conditions. The general experimental strategy follows the design of -Chapter 3 spatial validation experiments, whereby natural groupings of similar HRUs are identified using a k-means clustering algorithm. Within each cluster, the HRUs were split into M domains of equal area at random to generate the training and testing datasets for the ML algorithms. This random splitting ensures that the division is not influenced by geographic orientation within the cluster, which

could otherwise introduce biases related to topographic gradients or spatial patterns, Figure 4.2 provides a spatial representation of the scale hierarchy, showing a conceptual k-means cluster which is divided into M domains, each comprised of multiple HRUs. HRUs represent the fine-scale units, domains serve as the coarse computational units, and clusters are larger units that define similar regions for which ML weights are determined.

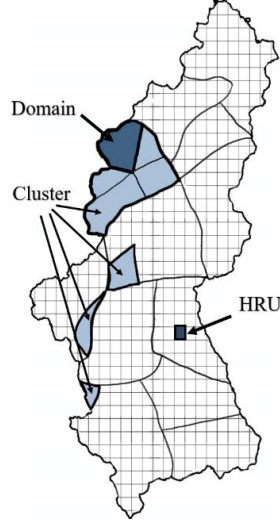


Figure 4.2: Provides a conceptual representation of the spatial subdivisions used in this study, showing the k-means (discontiguous) cluster divided into M (contiguous) domains, each composed of multiple HRUs.

The training dataset consists of M-1 domains with the remaining domain in the cluster used for testing. The ML algorithm is trained on 20 years of the forcing, state, and flux data from the training domains to estimate the fluxes for the testing domain over the entire study period. Table 4.1 provides a comprehensive overview of all the input and output variables included in the ML model. The simulated fluxes generated by the coarse resolution ML model are then integrated into a domain mass balance equation to compute SWE within each domain:

$$\overline{SWE}_i = \overline{SWE}_{i-1} + (\bar{P}_i - \bar{S}_i - \bar{M}_i) \cdot \Delta t, \quad (4.1)$$

where $\overline{SWE}_{i-1}$ [mm] is the computed domain mean SWE value from the previous timestep, $\bar{P}_i$ [mm/d] is the domain mean snowfall in the timestep i , $\bar{S}_i$ [mm/d] is the mean sublimation rate estimated by the ML algorithm, $\bar{M}_i$ [mm/d] is the mean snowmelt rate in time step i estimated by the algorithm, and Δt is the model time step, in days. The mass balance equation computes the domain-resolution SWE at a daily timestep, which is then compared against the spatially aggregated high-resolution Raven SWE for the same domain. Agreement between the aggregated fine-resolution model and the RF model, as evaluated using the KGE, demonstrates that, once trained, the upscaling methodology can accurately determine the fluxes of systems at the domain scale without the need for fine-resolution simulations.

Table 4.1: The contents of the input and output vector for the ML models where all mean forcings represent spatial averages at the specified computational scale. Terms in the input and output vectors marked with an asterisk denote variables that are optionally included in certain subsets of experiments.

Machine learning algorithms inputs and outputs		
	Forcing	State
Input variables	Spatial mean precipitation rate [mm/d]	Spatial mean SWE [mm]
	Spatial mean temperature [C]	Spatial snow-covered area [0...1] *
	Spatial mean wind speed [m/s]	
	Spatial mean net shortwave radiation [mJ/m ² /d]	
	Spatial mean net longwave radiation [mJ/m ² /d]	
	Spatial mean snowfall rate [mm/d]	
	Spatial mean snowfall temperature [C]	
	Spatial Standard Deviation of Forcing Variables *	
	Fluxes	States
Output variables	Spatial mean snowmelt rate [mm/d]	Spatial mean snow-covered area [0...1]
	Spatial mean sublimation rate [mm/d]	*

KGE is an error metric that evaluates the accuracy of hydrological models by considering correlation bias, and variability between simulated and observed data (Gupta et al., 2009). A KGE values of 1 indicates a perfect match between simulated and observed data, while values closer to 0 or negative indicate poorer model performance. Typically, a KGE value above 0.6 is considered adequate (Choi et al., 2021; Qi et al., 2021), while values below this threshold suggest that the model's performance may need improvement. For some evaluations, Mean Bias Error (MBE) and coefficient of determination (R^2) are used as additional error metrics. MBE measures the average difference between simulated ($\hat{\phi}$) and observed (ϕ) values, indicating the direction and magnitude of any systematic errors in the model. MBE is calculated as:

$$MBE = \frac{1}{n} \sum_{i=1}^n (\hat{\phi} - \phi) \quad (4.2)$$

where n is the total number of observations. An MBE value close to 0 is ideal, indicating minimal deviation, while large positive or negative values indicate a consistent over- or underestimation, respectively. R^2 , a popular error metric in ML, represents the proportion of variance in the observed data that is explained by the model, with values ranging from 0 to 1. R^2 is calculated as:

$$R^2 = 1 - \frac{\sum_{i=1}^n (O_i - S_i)^2}{\sum_{i=1}^n (O_i - \bar{O})^2} \quad (4.3)$$

where O_i is the mean observed values. An R^2 value close to 1 suggests that the model explains most of the variability in the data, indicating a good fit, while values closer to 0 indicate a weak fit.

An alternative method for evaluating experimental outputs involves comparing the mass balance estimated fluxes with fluxes naïvely upscaled through Raven, where naïve upscaling refers to applying small-scale process descriptions at larger scales while disregarding scaling effects and system heterogeneity (Ranjram, 2021). In this approach, the clusters generated in the experiment are converted into single computational units matching the extent of each domain. A simplified Raven model is then executed to simulate mean snowmelt, sublimation, and SWE for each of the single computational units using mean forcings and characteristics. The aggregated naïve upscaled SWE and fluxes is compared against the mass-balance computed coarse-resolution SWE and upscaled fluxes for each cluster.

4.2.5 Overview of Experiments

A series of experiments is here used to test the ML algorithm’s ability to capture relationships between fluxes, systems of states, and properties over a range of scales. This evaluation consists of four key experiments: (1) *ML Algorithm Comparison*, which broadens the scope of Burdett and Craig’s work (in review) by evaluating the performance of three ML algorithms rather than one, thereby identifying the ideal ML algorithm for the upscaling methodology; (2) *Spatial Resolution* experiments, which test the ML upscaling methodology across a range of scales by progressively increasing the number (M) of domains, as Chapter 3 only considered five domains, potentially limiting the scale dependency insights; (3) *Variable Reduction*, an approach to deepen understanding of the causative relationships between forcing variables and snow fluxes and states, where variables with low feature importance are iteratively removed, identifying the minimal number necessary to produce KGE values comparable to the full-variable model. In contrast to Chapter 3, where the three most significant variables were selected based on consistency across all clusters, this experiment provides insights into the variables most critical for each cluster individually; and (4) *Transferability of ML Relationships*, evaluating the applicability of relationships derived from clusters in the Kananaskis River Basin to the Oldman River basin. Together, these experiments critically examine the ML upscaling methodology and the constitutive relations it implicitly derives, aiming to refine and validate the approach across algorithmic, spatial, and variable dimensions.

4.2.6 Machine Learning Algorithm Comparison

In Chapter 3, only one ML algorithm (RF) was used. In the *ML Algorithm Comparison* (experiment 1), three ML algorithms RF, XGBoost, and ANN are compared to identify the most effective model for the ML upscaling methodology. Each of these algorithms was built with the Scikit-learn Python package and have not only been successfully applied in hydrologic research but also extensively compared

to one another in hydrologic literature (e.g., Ren et al., 2024; Yang et al., 2024). The success of these algorithms is known to depend on the specific study conditions. Therefore, selecting the most suitable ML algorithm is crucial for optimally configuring the upscaling methodology.

The RF model is an unsupervised, non-parametric algorithm within the decision tree family that comprises a combination of statistically independent trees to forecast classification and regression tasks (Breiman, 2001). In this method, the final simulated outputs from multiple decision trees are aggregated to produce the RF model's output. This ensemble approach improves the model's accuracy and enhances its robustness against overfitting. RF models are often preferred for streamflow forecasting based on their capability to handle large datasets with higher dimensionality (Lin et al., 2017). The use of RF algorithms allows one to understand the impacts of driving factors on response variables by evaluating the mean decrease in node impurity, also known as the Gini feature importance index. The Gini feature importance index measures the average reduction in residual error each time a predictor is selected to make a split during training. This is based on the principle that a binary split is made only when the residual error (or impurity) in the two descendant nodes is lower than in their parent node. The Gini index for a predictor is calculated by summing all gains across all trees and dividing by the number of trees, with a higher Gini index score indicating a more significant role for that feature within the decision trees (Mushtaq et al., 2024).

XGBoost is a decision tree algorithm based on the gradient boosting ensemble learning algorithm proposed by Chen et al. (2016). Gradient boosting is an ensemble learning algorithm that builds a series of decision trees sequentially, where each tree corrects the errors of the previous ones by minimizing a specified loss function. The iterative process results in a predictive model that combines the strengths of multiple weak learners to achieve high accuracy (Ren et al., 2024). XGBoost has been used in different fields of hydrology, including rainfall-runoff modelling (Li et al., 2024), forecasting ice phenomena (Graf et al., 2022), and stream temperature prediction (Weierbach et al., 2022). Similar to RF, the interpretability of XGBoost stems from its ability to estimate the Gini feature importance index.

Although not generally considered an interpretable ML method, ANNs are included in this study due to their high success and frequent comparison to other ML algorithms in hydrology. An ANN is a data processing system that consists of a highly interconnected network of simple processing elements called neurons. In an ANN, neurons are arranged in layers with connections between adjacent layers, and the strength of these connections is referred to as the "weight" (Tanty and Desmukh, 2015). The ANN connection weights of nodes in the hidden layers are adjusted during training to obtain the best model. It has strong fault tolerance and learning ability (Lin et al., 2023). Despite the challenge of interpreting ANNs due to their complexity and opacity, methods like SHAP can be applied to enhance interpretability. SHAP works by assigning each feature a Shapley value, representing its contribution to the prediction by analyzing the impact of including or excluding it; this approach is model agnostic, making it applicable to any algorithm.

The *ML Algorithm Comparison* experiment applied the three ML algorithms with 12 clusters in the Kananaskis River Basin to estimate the snow-related fluxes necessary for the mass balance equation to compute SWE. Each cluster is divided into five domains, four of the domains are used as a training dataset and the remaining domain as a testing dataset. The architecture of the RF model replicates the structure used in Chapter 3. The XGBoost and ANN models underwent a training and validation period (1994–2000) during which hyperparameters were adjusted iteratively using a range of incremental values that were manually selected, following a grid-search style approach similar to the hyperparameter tuning frameworks available in scikit-learn Python package. The final architectures of these ML algorithms are detailed in Table 4.2. To mitigate potential information loss due to aggregation, the ML models training dataset include the distributional information of snow temperature, temperature, and shortwave radiation. These forcing variables were selected because they consistently emerged as the most significant across all clusters in the RF, XGBoost, and ANN models, as identified by SHAP during the training and validation period. To assess the performance of the ML models, the upscaled snowmelt, sublimation, and mass-balance estimated SWE values are compared against the aggregated high-resolution Raven SWE and the naïvely upscaled model through the KGE, MBE, and R^2 error metrics.

Table 4.2: An overview of the ML algorithm comparison setup, outlining the architectures of each ML model, clustering configuration, domain specifications, included distributional information, and the error metrics used for evaluation.

Machine learning algorithm comparison			
Model architecture	Random Forest (RF)	XGBoost	Artificial Neural Network (ANN)
	Number of estimators: 6 Max depth: 12 Min samples split: 5	Number of estimators: 125 Max depth: 20 Min samples split: 5 Learning rate: 0.05	Hidden layer sizes: 120, 60 (2 hidden layers) Max iterations: 500 Learning rate: 0.05
Number of clusters	12		
Number of domains	5		
Distributional Information	Yes, snow temperature, temperature, and shortwave radiation		
Evaluation criteria	KGE, MBE, R^2 , naïve model		

4.2.7 Spatial Resolution Comparison

The *Spatial Resolution* experiment (experiment 2) aims to identify the optimal domain resolution for applying upscaling relationships within each cluster, as only five domains were originally used in Chapter 3. This experiment seeks to determine the finest and coarsest resolution at which these relationships remain sufficiently accurate, defined as the point where the upscaled snowmelt, sublimation, and mass-balance derived SWE values maintain strong agreement with fine-resolution Raven values, as measured by KGE. Choosing the right number of domains is critical: too few domains may lead to excessive generalization, while too many may result in minimal variability within each cluster (Liu et al., 2020). To identify the ideal resolution, the number of domains within each of the 12 clusters was systematically adjusted, beginning with 4 and doubling incrementally up to 64 (e.g., 4, 8, 16, 32, 64). For each domain configuration, the general experimental design was applied using an RF model, where 25% of the domains were reserved for testing, and the remaining 75% of domains were used for training across the entire study period, as shown in Table 4.3. Each configuration's performance was assessed by comparing upscaled snowmelt, sublimation, and mass-balance-derived SWE values with fine-resolution SWE using KGE. Distributional information was excluded from this experiment.

Table 4.3: Outlines the setup for the spatial resolution comparison experiment using an RF model.

Spatial resolution comparison										
Model architecture	Random Forest (RF) Number of estimators: 6, max depth: 12, min samples split: 5									
Number of domains	4		8		16		32		64	
Domain Training/Testing Split	<i>Train</i> 3	<i>Test</i> 1	<i>Train</i> 6	<i>Test</i> 2	<i>Train</i> 12	<i>Test</i> 4	<i>Train</i> 24	<i>Test</i> 8	<i>Train</i> 48	<i>Test</i> 16
Number of clusters	12									
Distributional Information	No									
Evaluation criteria	KGE									

4.2.8 Variable Reduction

The *Variable Reduction* experiment (experiment 3) employs a Leave-One-Out (LOO) analysis to identify the smallest set of forcing variables needed to achieve performance comparable to the full-variable ML model. This full-variable model consists of a RF model using 5 domains (4 for training and 1

for testing), including distributional information. The 5-domain split was chosen based on results from the *Spatial Resolution* experiment, fitting within the identified optimal range for upscaling resolution while also aligning with the *ML Model Comparison* experiment spatial resolution. This ensures sufficient variability for model learning while preventing excessive generalization or loss of detail. The LOO analysis begins by running the RF model with all forcing variables included. The variables are then ranked according to their Gini feature importance scores, and the variable with the lowest score is removed. This process is repeated iteratively, removing one variable at a time, until only a single variable remains. At each iteration, KGE values for the ML upscaling mass-balance-estimated SWE, snowmelt, and sublimation are compared to those of the full-variable model. For each cluster, the variable combination with the fewest variables but comparable KGE values to the full-variable model is selected as the final output of the LOO analysis. The final LOO models are then rerun, and their outputs are compared to the full-variable ML model and aggregated high-resolution Raven model to validate performance.

Through this methodical reduction, the LOO analysis helps pinpoint the minimal number of variables necessary for the model to maintain its predictive accuracy, thus optimizing efficiency without compromising performance. This approach not only validates the significant variables identified through the ML model feature importance indexes but also may provide insights into which variables most and least significantly impact system fluxes at the scale of interest. By identifying the essential variables, we can relate the ML-derived understanding of small-scale relationships to established knowledge of snow ablation processes and their associations with meteorologic, topographic, and landcover characteristics documented in the literature. This enables us to evaluate how accurately these relationships are represented in the upscaled model.

Table 4.4: Overview of the Variable Reduction experiment outlining the ML model used, the number of clusters and domains, the inclusion of distributional information, and evaluation criteria that are used.

Variable Reduction	
Model architecture	Random Forest (RF) Number of estimators: 6, max depth: 12, min samples split: 5
Number of clusters	12
Number of domains	5
Evaluation criteria	KGE
All Input Forcing Variables	Spatial mean precipitation rate [mm/d], temperature [C], wind speed [m/s], net shortwave radiation [mJ/m ² /d], net longwave radiation [mJ/m ² /d], snowfall rate [mm/d], snowfall temperature [C], SWE [mm], snow-covered area [0...1], and spatial standard deviation of all variables

4.2.9 Transferability of ML Relationships

For the *Transferability of ML Relationships* experiment (experiment 4), the upscaling relationships derived from the *ML Algorithm Comparison* experiment 5-domain RF model for the Kananaskis model are applied to the Oldman River model. The *Transferability of ML Relationships* experiment is a spatial validation assessment that evaluates whether these relationships are transferable among study areas that share similar topographic and landcover characteristics. The Oldman River model was developed in collaboration with the Alberta Innovates project and inspired by the Chernos et al. (2022) Oldman River model, which is an adapted version of the distributed HBV-EC hydrological model (Bergström, 1992), emulated within the Raven framework. For the Oldman River model, subbasins are defined by extracting the Hydrologic Unit Code (HUC) Watersheds that intersect with the study area. A DEM layer with a 25m resolution, obtained through the R “elevatr” package (Hollister and Shah, 2017), combined with landcover data from Natural Resources Canada (NRCan) (Natural Resources Canada, 2015), was used for the delineation of HRUs. The Kananaskis model used the same DEM and landcover data sources for delineation. Additional processing was applied to delineate the Oldman HRUs to a finer resolution, resulting in approximately 30,000 HRUs, each with a resolution of $\sim 0.10 \text{ km}^2$. Although this resolution is coarser than that of the Kananaskis HRUs, it was selected for the Oldman model to manage computational demands effectively while retaining sufficient spatial detail.

While the Oldman model uses coarser resolution HRUs compared to the finer-resolution HRUs in the Kananaskis model, both models were developed using the same DEM and landcover datasets, ensuring a consistent foundation for both discretization. Although some scale-sensitive variables such as slope and aspect may be less precisely represented in the coarser Oldman HRUs, the overall topographic and landcover gradients remain intact, allowing the model to generalize effectively across regions with similar physiographic characteristics. The Oldman River basin extends further into the lowlands compared to the Kananaskis basin, resulting in a higher proportion of low-elevation, relatively homogeneous shrubland and grassland areas. In such landscapes, a coarser resolution is likely sufficient to capture key spatial patterns in snow accumulation and melt, further supporting the applicability of the scaling relationships across both models. Although the Oldman River model is an adapted version of the distributed HBV-EC hydrological model, which typically represents snow accumulation and melt using a single-layer, degree-day based approach, the snowmelt routine was modified to more closely align with the snowmelt process used in the Kananaskis model.

After training the k-means model on the Kananaskis River Basin landcover and topographic characteristics, it is saved using Python's “pickle” library for future use (Van Rossum and Drake, 2009). This package allows for Scikit-Learn trained ML model's to be efficiently reloaded and applied to new datasets without the need for retraining. The k-means model is then reloaded and applied to the Oldman River HRUs, assigning them to the predefined Kananaskis River Basin clusters. This ensures consistent classification of new data based on the established cluster structure. Similarly, the RF model from the *ML Algorithm Comparison* experiment is also stored using the “pickle” library. For each cluster, the

corresponding stored RF model is reloaded and uses coarse resolution forcing and state variables from the Oldman model as inputs to estimate snowmelt, sublimation, and SCA. The RF-estimated snow fluxes are then integrated into the mass balance equation to calculate SWE. To assess the performance of the ML relationship transfer, the ML-upscaled estimates for snowmelt, sublimation, SCA, and SWE are compared against outputs from both the aggregated high-resolution Raven model and the naïvely upscaled Oldman River model using KGE, MBE, and R^2 error metrics.

Table 4.5: Overview of the Transferability of ML Relationships experiment outlining the ML model used, the number of clusters and domains, the inclusion of distributional information, and evaluation criteria that are used.

Transferability of ML Relationships	
Model architecture	Random Forest (RF) Number of estimators: 6, max depth: 12, min samples split: 5
Number of clusters	12
Number of domains	5
Distributional Information	Yes, snow temperature, temperature, and shortwave radiation
Evaluation criteria	KGE, MBE, R^2 , naïve model

4.3 Results and Discussion

4.3.1 Machine Learning Algorithm Comparison

In this study, three ML algorithms, RF, XGBoost, and ANN, are compared to identify the most effective model for the ML upscaling methodology in deriving constitutive relations for snow-related fluxes. Following the steps previously outlined, each algorithm was tested on 12 clusters within the Kananaskis River Basin (Figure 4.3). The clusters were divided into five sections, with four used for training and the remaining one section for testing. The three algorithms were trained on daily mean fluxes from the training sections over a 20-year period and then used to estimate the daily mean fluxes for the testing domain over the entire study period. The simulated fluxes from the coarse resolution ML models are then integrated into a domain mass balance equation to compute SWE, demonstrating how the upscaling methodology would function within a complete model.

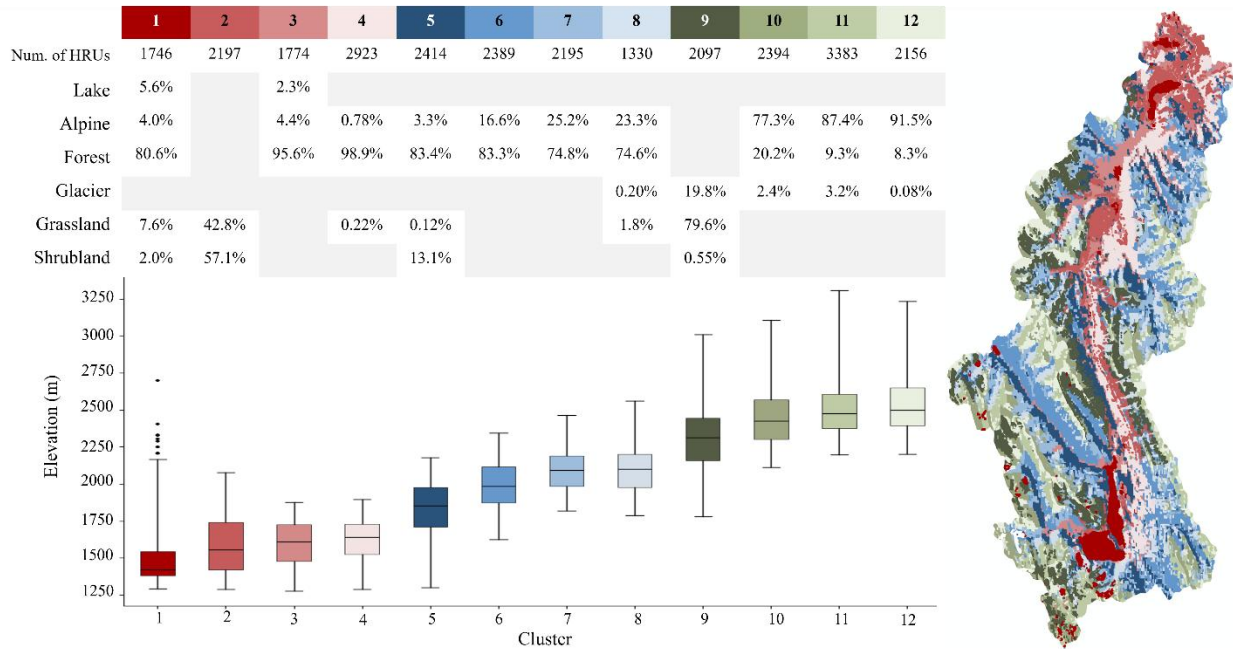


Figure 4.3: A plot of the 12 k-means clusters of the Kananaskis River Basin and a corresponding table indicating the number of HRUs in each section, elevation, and landcover percentages in each cluster.

The predictive performance of each ML algorithm in the *ML Algorithm Comparison* experiment (experiment 1), evaluated using KGE, MBE, and R^2 , is shown in Figure 4.4. Among the evaluated ML models, RF consistently achieved higher KGE and R^2 values across snowmelt, sublimation, SCA, and mass-balance estimated SWE. This indicates a strong agreement between the upscaled values and aggregated high-resolution Raven estimates. The robust performance of RF is further confirmed by its near-zero MBE across all variables, suggesting it neither systemically overestimates nor underestimates predictions. XGBoost also performs well, with KGE and R^2 values comparable to those of RF. However, XGBoost shows higher variability in MBE, particularly for mass-balance estimated SWE and snowmelt, indicating a tendency to overestimate the upscaled values. In contrast, the ANN tends to underperform, displaying lower and more variable KGE and R^2 values particularly for snowmelt and sublimation. For SWE, the ANN model exhibits significant negative MBE's, making it prone to underestimation of mass-balance estimated SWE values. For sublimation and SCA, all ML algorithms exhibit near-zero MBE, indicating that none of the models systematically skew predictions for these variables.

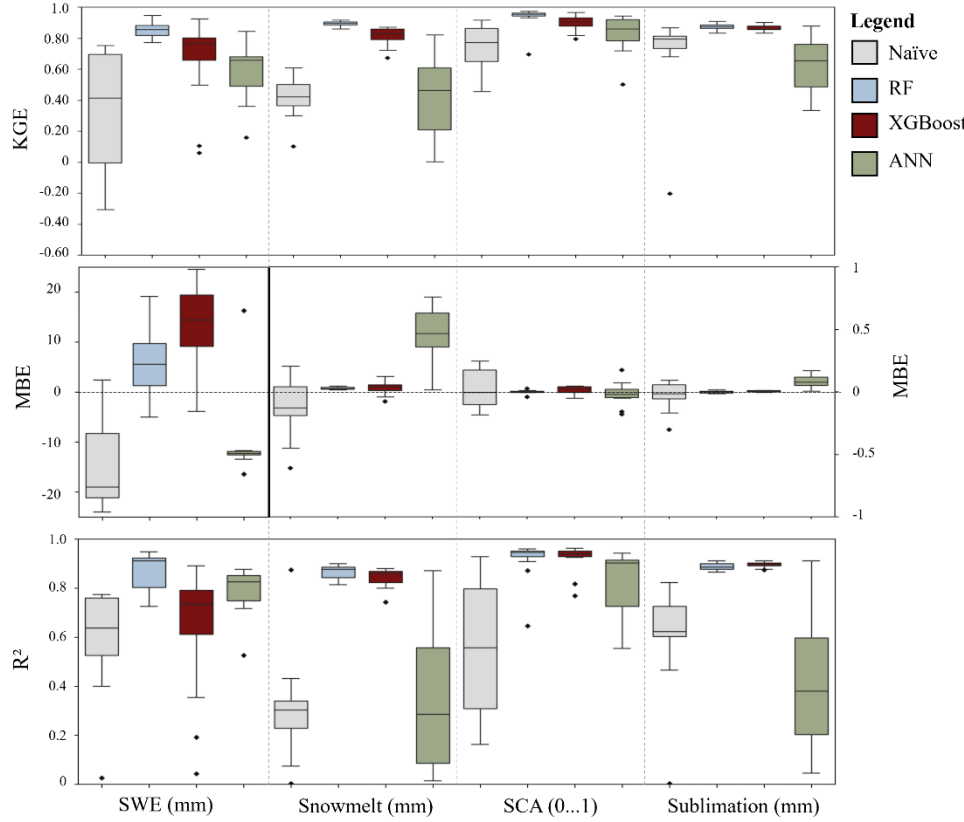


Figure 4.4: Illustrates the KGE, MBE, and R^2 values for all clusters, evaluating the agreement across various variables, including Raven naïvely upscaled, RF, XGBoost, and ANN-estimated snowmelt, sublimation, SCA, and mass balance estimated SWE agreement with the aggregated high-resolution Raven estimates.

The naïve upscaling model snow accumulation and ablation estimates demonstrated lower KGE and R^2 values across all variables, except for sublimation, where it outperforms the ANN model. The naïve Raven model shows higher variability in MBE values, indicating a relative tendency to underestimate mean SWE and snowmelt, while slightly overestimating mean SCA. The performance of the naïve model underscores that standard upscaling methodologies (which disregard the influence of sub-computational unit heterogeneity) often overlook the impact of scaling effects. This oversight impacts the performance as it may fail to account for the effects of varying topographic and landcover characteristics.

Tree-based ML algorithms, such as RF and XGBoost, are widely recognized in the literature as highly effective for modelling snow distribution based on topographic and vegetation drivers (e.g., Bennett et al., 2022; Dharmadasa et al., 2022; Litaor et al., 2008; Meloche et al., 2022; Moreno et al., 2010; Revuelto et al., 2020; Winstral et al., 2002). The high performance of tree-based methods can be attributed to their robust ability to capture non-linear relationships and manage data variability. In contrast, the lower

performance of ANNs in ML upscaling for snow modelling is anticipated due to their sensitivity to parameter tuning and their challenges in handling data irregularities. Snow modelling datasets often include abundant zero values, which represent conditions of no snow cover, no snowmelt, or no sublimation. ANNs can be particularly sensitive to this noisy, zero-dominated data, as their optimization is heavily influenced by the most common values in the dataset. This can lead ANNs to overfit on these frequent zero values, reducing their capacity to accurately predict non-zero events. While XGBoost generally demonstrates greater robustness than ANNs, it may also struggle with zero-dominated datasets. Since XGBoost builds trees sequentially by correcting errors from prior trees, it may inadvertently amplify the influence of noise and zero values, resulting in bias and reduced performance. RF, on the other hand, seems to mitigate these imbalances more effectively due to its ensemble structure, which aggregates predictions from multiple trees that sample features and observations independently. This approach distributes the impact of zeros and noise across the model, yielding more balanced and accurate predictions overall.

4.3.2 Spatial Resolution Comparison

The *Spatial Resolution* experiment (experiment 2) aims to identify the optimal computational resolution for applying upscaling relationships within clusters. To find the ideal resolution, the number of domains within each of the 12 clusters was systematically adjusted, beginning with 4 domains, and doubling incrementally up to 64 (leading to configurations with 4, 8, 16, 32, and 64 domains). Given the strong performance of RF models in the *ML Algorithm Comparison* experiment, RF was selected as the ML algorithm for the remainder of this study. For each cluster, 25% of the domains were randomly selected as the testing dataset using Python's random sampling functions, while the remaining 75% of domains were assigned to the training dataset. This random split was performed independently within each cluster. Although the selection was random, the domains within each cluster tend to be spatially contiguous and adjacent, rather than widely dispersed across space. The RF algorithm is then trained on the daily mean fluxes generated by the fine scale model of the training domains over a 20-year period to estimate the daily mean fluxes of the testing section for the study duration.

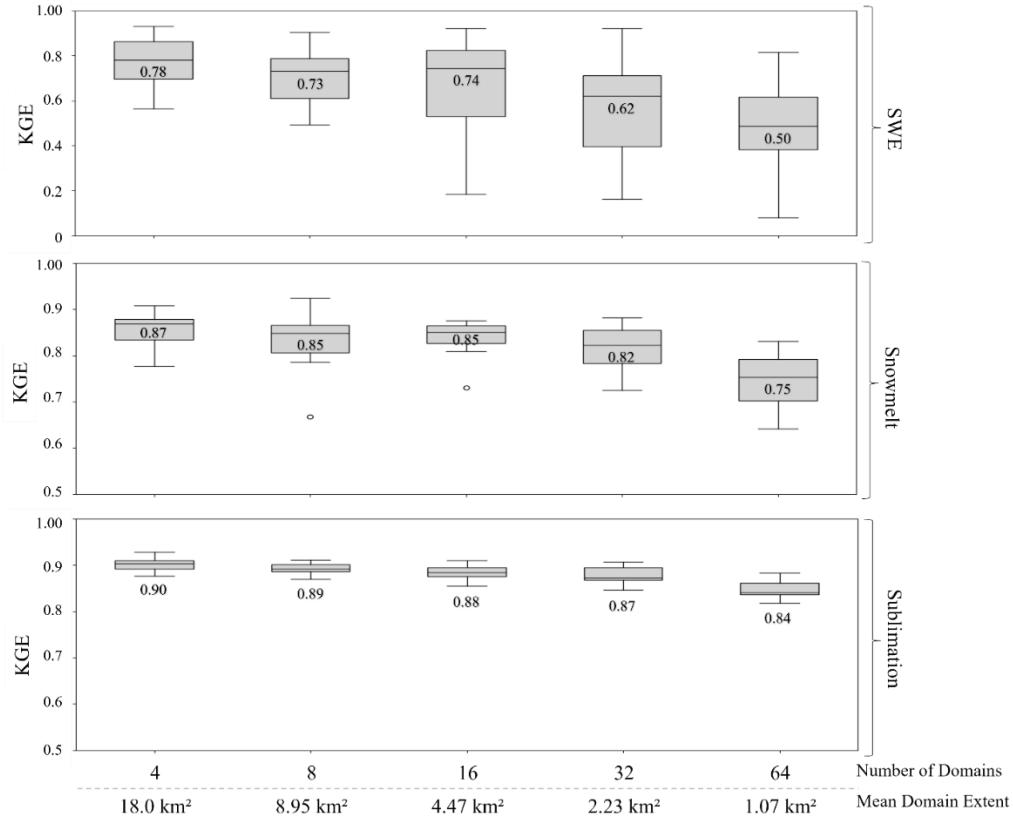


Figure 4.5: KGE values for mass-balance estimated SWE, snowmelt, and sublimation compared to fine-resolution Raven model snow-related fluxes as the number of domains within each of the 12 clusters was systematically adjusted (4, 8, 16, 32, and 64).

The *Spatial Resolution* experiment results are presented in Figure 4.5, depicting box plots of the KGE values for the mass-balance estimated SWE, snowmelt, and sublimation compared to fine-resolution Raven model snow-related fluxes. The number of domains corresponds to the scale of the spatial aggregation: the 4-domain model divides each cluster into 4 domains with a mean size of 18.0 km², while the 64-domain model creates domains with a mean size of 1.07 km². It is expected that the coarse aggregation of the 4-domain case will lead to decreased performance due to the higher degree of variability being homogenized. On the other hand, there may be diminishing returns for a cluster divided into many domains as similar domains are likely to yield similar training data, limiting the model's exposure to meaningful variability.

For SWE, the RF model performed well with 4 to 8 domains, yielding median KGE values of 0.78 and 0.73, comparable to the 5-domain resolution used in the previous *ML Algorithm Comparison* experiment. At 16 domains, the range of SWE KGE values increased, but the median KGE remained stable at 0.74. As the number of domains increased to 16, median SWE KGE values remained consistent but began to exhibit a wider range of KGE values. Further increases to 32 and 64 domains resulted in reduced performance; while the median KGE values remained at or above 0.5, the wider range of KGE values and

the presence of outliers indicated reduced model accuracy. Snowmelt followed a similar trend, with stable performance up to 16 domains and declines at 32 and 64. In contrast, sublimation KGE values remained high with minimal outliers across all spatial resolutions, demonstrating the model’s robustness in estimating sublimation fluxes across varying computational resolutions.

Increasing the number of domains expands the training dataset: the model trains on N years of data from 3 domains in the 4-domain model, compared to a 64-domain setup where it trains on 48 domains, ultimately increasing sample diversity. However, smaller domains also tend to capture more noise rather than generalizable patterns, likely contributing to performance degradation with an increased number of domains. This results in greater variability in model performance and an increase in outliers, as evidenced by the wider interquartile range and lower KGE values for 64 domains. Similar sensitivity to spatial resolution has also been observed in process-based snow models, where varying spatial scales affect SWE and SCA estimates across heterogeneous landscapes, with data granularity improving model training only up to an optimal level. For instance, Barnhart et al. (2024) found that coarser model resolutions (500–1,000 m) were sufficient for streamflow correlations in larger watersheds, while finer model resolutions (30–150 m) provided more benefit in small, complex terrains. These results highlight the importance of selecting a computational resolution that preserves data diversity without compromising model robustness. Additionally, the *Spatial Resolution* experiment emphasizes the need to balance accuracy with computational efficiency, as higher domain counts increase computational costs for training without corresponding gains in model performance.

The performance of the RF models was initially expected to improve as the number of domains increased to 32 and 64, based on the assumption that smaller domain extents, being closer in size to the original high-resolution model, would enhance model accuracy. Instead, the opposite trend emerged. In these domains 32 and 64, the model had fewer internal variations to learn from, and each domain provides only limited heterogeneity, essentially mimicking single-HRU behaviour. As a result, the model may be more prone to overfitting localized noisy inputs rather than capturing generalizable patterns.

4.3.3 Variable Reduction

The *Variable Reduction* experiment (experiment 3) employs a LOO approach to identify the smallest set of variables required for the RF model to perform comparably to the full-variable model. The intent of this experiment is to determine the key controls on the ML-estimated upscaling relationships by systematically removing less influential predictors and assessing model performance. The experiment begins by running the RF model with all forcing variables including the distributional information of each forcing and state variable. In each iteration, the variable with the lowest Gini feature importance index score was removed from the training dataset. This process continues until only one variable remains in the training dataset, with KGE values from each iteration compared to identify the smallest variable set

that performs similarly to the full-variable model. Once the minimum set of variables is identified, their final KGE values are compared to those of the full model and its feature importance index. These key variables are then cross-referenced with established knowledge of snow ablation and accumulation processes from the literature to further assess their impact on scaling. This approach enables a data-driven understanding of the dominant controls on upscaling snow processes, helping to refine model complexity while preserving predictive skill.

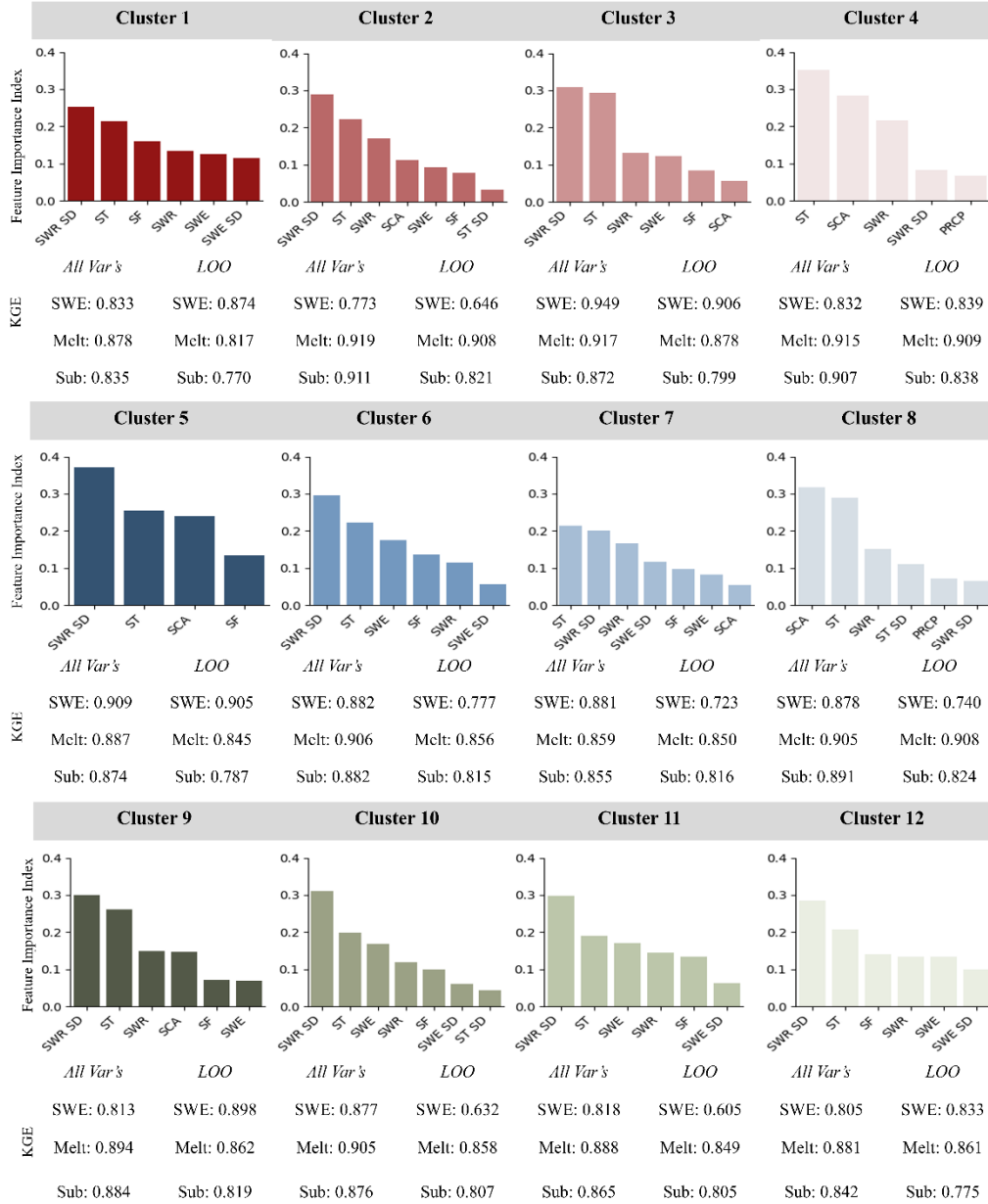


Figure 4.6: A plot displaying the significant features identified by the LOO method and KGE values for both the full-variable and LOO method RF models across each cluster. Acronyms: SWR-SD (Shortwave Radiation Standard Deviation), ST (Surface Temperature), SWR (Shortwave Radiation), SCA (Snow-

Covered Area), SF (Snowfall), SWE (Snow Water Equivalent), SWE-SD (Snow Water Equivalent Standard Deviation), PRCP (Precipitation), Wind (Wind Speed), Temp (Temperature).

In the LOO experiment, all clusters achieved a KGE value greater than 0.7 for mass-balance estimated SWE and sublimation, demonstrating a moderate to strong level of agreement between the RF all-variable model outputs and the Raven fine resolution model, as shown in Figure 4.6. Notably, KGE values for snowmelt were also consistently above 0.7 and were generally higher than those for sublimation, indicating stronger model performance for snowmelt. Despite the reduction in the number of training variables, snowmelt KGE values remained consistently similar to the all-variable model, with a maximum performance decrease of 0.13. The majority of clusters required only six variables to achieve KGE values comparable to the all-variable model. Among these, net shortwave radiation, SWE, SCA, and snow temperature were consistently the most significant variables in the feature importance index across all clusters. However, variable significance does not necessarily imply that the RF model is inferring meaningful physical relationships, as importance is independent of predictive skill. For instance, melt and sublimation fluxes should exhibit a linear dependence on SCA at the start of each time step and cease entirely when SWE reaches zero. It is important to note that SWE and SCA inputs serve as initial conditions for flux estimation, not as training data, as they are carried forward from the previous timestep to define the snowpack state at the start of the current timestep. In the two-layer snow balance model used in the fine scale model (Wigmosta et al., 1994), snow temperature and incoming energy play a pivotal role in simulating snow accumulation and ablation. Snow temperature determines when snow begins melting within the two-layer snowmelt model while short and longwave radiation provide most of the incoming energy for melt (Golding & Swanson, 1986). Snow temperature is included as an input forcing variable so that the RF model is more capable of capturing the energy balance dynamics governing snowmelt. Snow temperatures feature importance ranking likely reflects its role in controlling phase transitions and regulating energy availability for snowmelt.

In the LOO experiment, longwave radiation and humidity were not retained as significant variables in any of the clusters. Longwave radiation was previously identified as a consistently significant variable across all clusters in Chapter 3. The difference in variable importance between the RF models used in the previous and current studies may be attributed to the increased landcover complexity in the current Raven model. The Burdett and Craig model used a simplified land cover classification with only forested, non-forested, and lake categories as a proof of concept. In contrast, the Raven model in this study used detailed NRCan landcover data, matching the source used by Chernos et al. (2022) for the Oldman River Basin. This increase in landcover complexity highlights the impact of detailed vegetation and landcover information on the model's response to snow dynamics. The reduced significance of humidity suggests that snow ablation fluxes were primarily constrained by energy availability rather than moisture gradients. While relative humidity plays a key role in sublimation by quantifying the vapor

pressure deficit, it may have been overlooked by the RF model, which estimated both snowmelt and sublimation. Since the model struggled more with estimating sublimation than snowmelt, it likely prioritized radiation and temperature variables, as snowmelt is predominantly driven by energy balance. However, if the RF model were solely focused on estimating sublimation, there would likely be a shift in variable significance, with humidity perhaps playing a more prominent role due to its direct influence on vapor pressure deficits and sublimation rates.

The standard deviation of all variables was included as additional inputs to the RF model, with the assumption that it would account for some sub-grid variability impacts. Across all clusters, the standard deviation of shortwave radiation was a predominantly significant variable in the RF feature importance index. The inclusion of the standard deviation of shortwave radiation therefore appears to improve the model's ability to capture the distribution of energy within the system and compensate for information loss during aggregation. Given that net shortwave radiation is often the primary incoming energy source for snow ablation (Golding and Swanson, 1986), it is unsurprising that accounting for its variability enhanced the accuracy of estimated fluxes and state variables. Similarly, the distributional information of SWE and snow temperature ranked highly in the feature importance index across most clusters in the reduced-variable model, as these variables directly influence water availability and the energy balance within the snowpack.

In mountainous terrain, snow accumulation patterns are primarily controlled by elevation, which influences orographic precipitation and leads to lower air temperatures, creating conditions that support persistent snowpack throughout the year (McKay and Gray, 1981; Anderton et al., 2004). Higher elevations are often dominated by alpine landcover, as reflected in the highest elevation clusters 10, 11, and 12. These clusters have minimal forest and glacier cover and are characterized by steep slopes. At high elevations, the exposed nature of alpine areas means that wind plays a more critical role in snowpack dynamics. Wind can influence snowmelt through enhanced turbulent heat transfer, and strong winds can erode snow, exposing bare ground that accelerates melt. Additionally, wind promotes sublimation by continuously replacing the air above the snow surface with dry air, facilitating the direct conversion of snow to vapor (Liston and Sturm, 1998). In contrast, wind speed is likely less influential in clusters with higher forest cover, where the canopy reduces snow transport and sublimation by sheltering the snowpack from wind. This trend is evident in the LOO analysis, as cluster 10, with a higher percentage of forest cover, shows reduced sensitivity to wind speed compared to clusters 11 and 12. While windspeed does not have the highest variable importance in clusters 11 and 12, its exclusion from the training dataset results in a KGE loss of approximately 0.1, indicating its impactful role in these exposed, alpine environments. This highlights the value of including windspeed when deriving scaling relationships for snow ablation processes in clusters where the open terrain allows wind-driven processes to play a prominent role.

Cluster 5 needed only four variables: standard deviation of net shortwave radiation, snow temperature, SCA, and snowfall to achieve KGE values for SWE comparable to those obtained with the

full variable model. However, performance slightly declined for snowmelt and more noticeably for sublimation, with the KGE decreasing by 0.09. This cluster is located at mid-elevation and features variable landcover, including grassland, shrubland, and alpine terrain, though it is predominantly forested. The significance of distributional information for net shortwave radiation is likely attributed to the variable landcover, where canopy shading reduces direct solar energy reaching the snowpack, effectively moderating melt rates. Landcover further promotes ablation in alpine and forested terrain through interception, where snow captured on branches is exposed to wind and radiation, increasing sublimation losses before the snow reaches the ground and affecting SCA (Pomeroy et al., 1998). This highlights how a simplified model was sufficient to achieve KGE values comparable to the full-variable model. While adding variables like wind speed may provide some improvements in specific cases, the results suggest that a core set of key variables can adequately capture the snowpack dynamics.

4.3.4 Relationship Transfer

The *Relationship Transfer* experiment (experiment 4) assesses the transferability of RF-derived constitutive relationships between study areas that share similar topographic and landcover characteristics. The trained k-means clustering model from the Kananaskis River Basin is applied to the Oldman River Basin HRUs, assigning these HRUs to the predefined Kananaskis clusters. Figure 4.7 shows the spatial distribution of the clusters, number of HRUs, variability in elevation, and landcover percentages within the Oldman River basin. The trained RF model for each cluster from the 5-domain *ML Algorithm Comparison* experiment, which include the distributional information of snow temperature, net shortwave radiation, and SWE, are applied to the corresponding Oldman clusters. Coarse resolution forcing and state variables from the Oldman model serve as inputs to estimate snowmelt, sublimation, and SCA. The RF-estimated snow fluxes are incorporated into a mass balance equation to estimate SWE. To evaluate the performance of the ML relationship transfer, the ML-upscaled values for snowmelt, sublimation, SCA, and mass-balance estimated SWE are compared against those from a fine-resolution Raven model of the Oldman basin and a naïvely upscaled Oldman River model. This experiment provides insight into the robustness and adaptability of ML-derived snow constitutive relationships across regions that share similar characteristics, highlighting the potential of ML derived upscaling relationships to

generalize beyond the boundaries of the original training environment

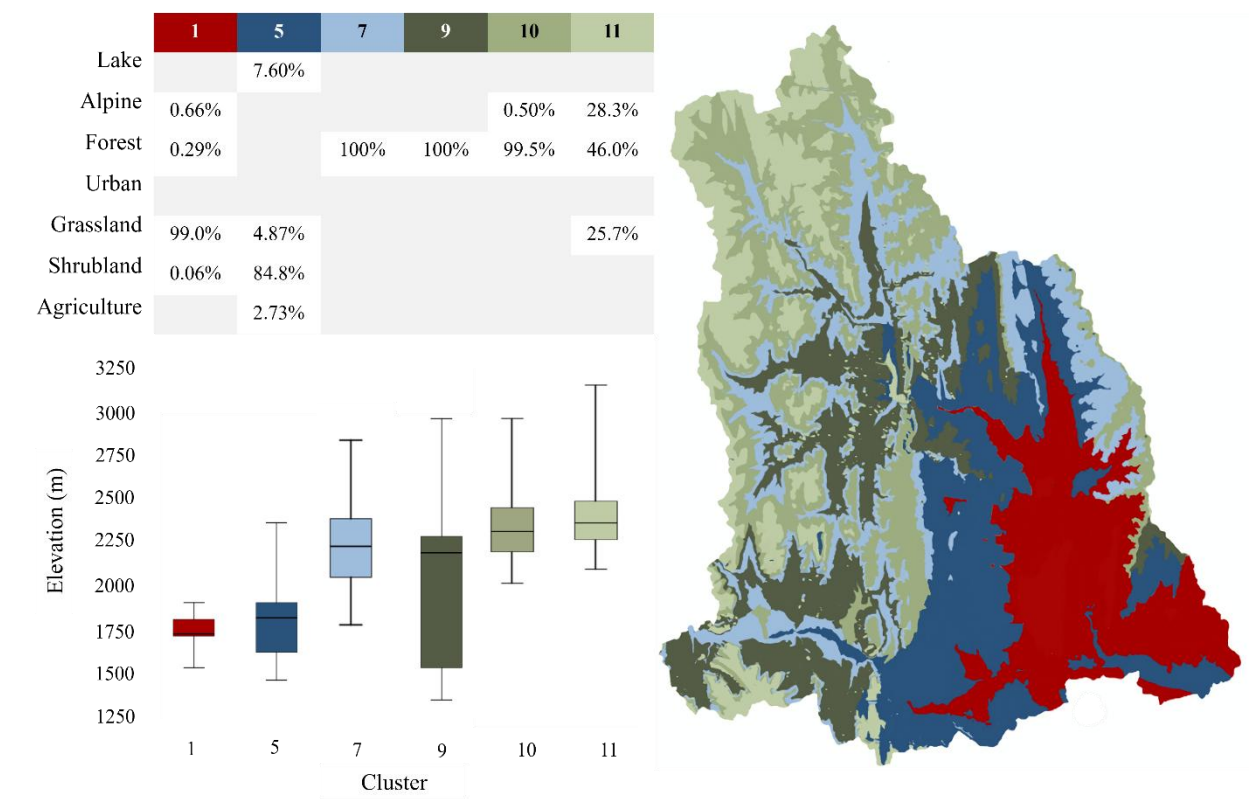


Figure 4.7: A plot of the k-means clusters of the Oldman River Basin and a corresponding table indicating the number of HRUs in each section, elevation, and landcover percentages in each cluster.

Several clusters derived in the Kananaskis analysis are notably absent in the Oldman Basin, likely due to Oldman’s distinct landcover features, such as the presence of glaciers in Kananaskis but not in Oldman, and agriculture and urban areas found in Oldman but not Kananaskis. The landcover distribution between these basins also varies: the Oldman River Basin has a greater proportion of shrubland (14%) and grasslands (22%), while in Kananaskis, shrubland comprises only 1.6% and grasslands 5% of the total area. Nevertheless, both basins have similar forest coverage, each around 60%. Additionally, both study areas exhibit similar topographic characteristics, as reflected in the clustering analysis: each cluster spans comparable elevation ranges across the two basins, with cluster 11 consistently appearing at higher elevations and cluster 5 predominantly at lower elevations. Both basins also share a comparable distribution of north- and south-facing slopes. These shared features make the study areas well-suited for evaluating the transferability of ML-based constitutive relationships between basins.

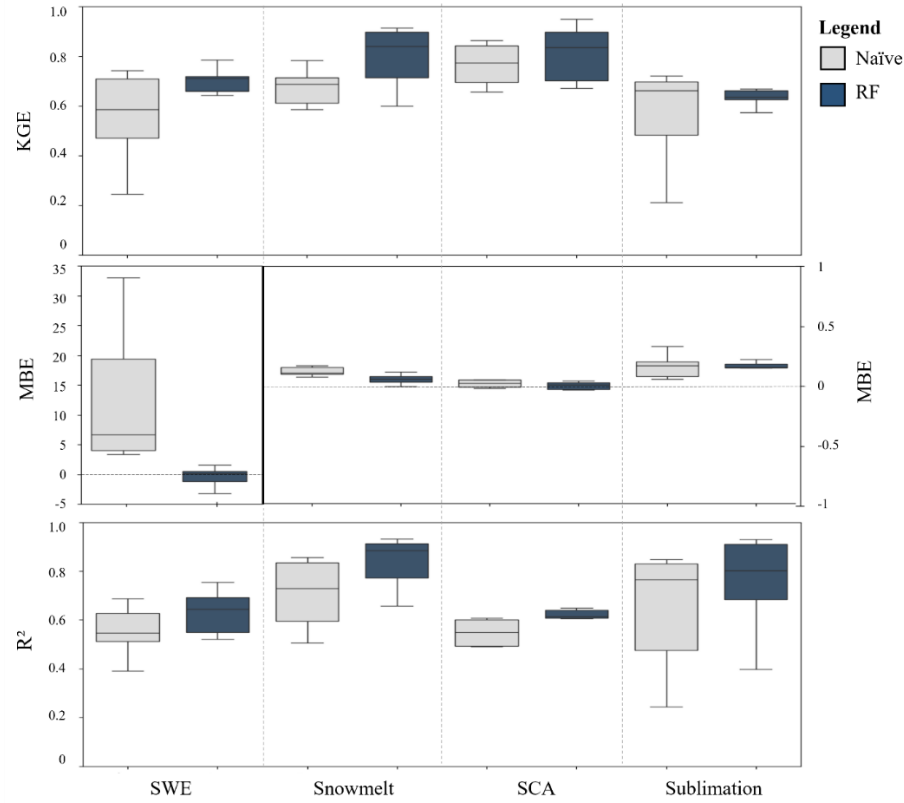


Figure 4.8: KGE, MBE, and R^2 values for all clusters, evaluating the agreement across various variables, including Raven’s naïvely upscaled estimates and the RF-transferred relationships from Kananaskis to the Oldman River Basin.

The RF model’s predictive performance for the *Relationship Transfer* experiment is shown in Figure 4.8. In each Oldman Basin cluster, the RF model achieved KGE values exceeding 0.7 for mass-balance estimated SWE, indicating moderate model agreement. The RF model’s mass-balance SWE had MBE values close to zero, further demonstrating accurate overall predictions with reduced tendency for skewed estimations. In contrast, the naïve upscaling approach yielded SWE estimations with a broader range of KGE and R^2 values with lower median values, suggesting a lower level of agreement for the naïve model. Additionally, the naïve model exhibited higher positive MBE, indicating a tendency to overestimate SWE. Figure 4.9 presents a time series comparison of Raven fine-resolution SWE, the Oldman mass balance estimated SWE, and the naïve approach for cluster 1. The plot highlights the naïve model’s tendency to overestimate SWE, particularly during peak accumulation periods, while the RF model underestimates SWE values.

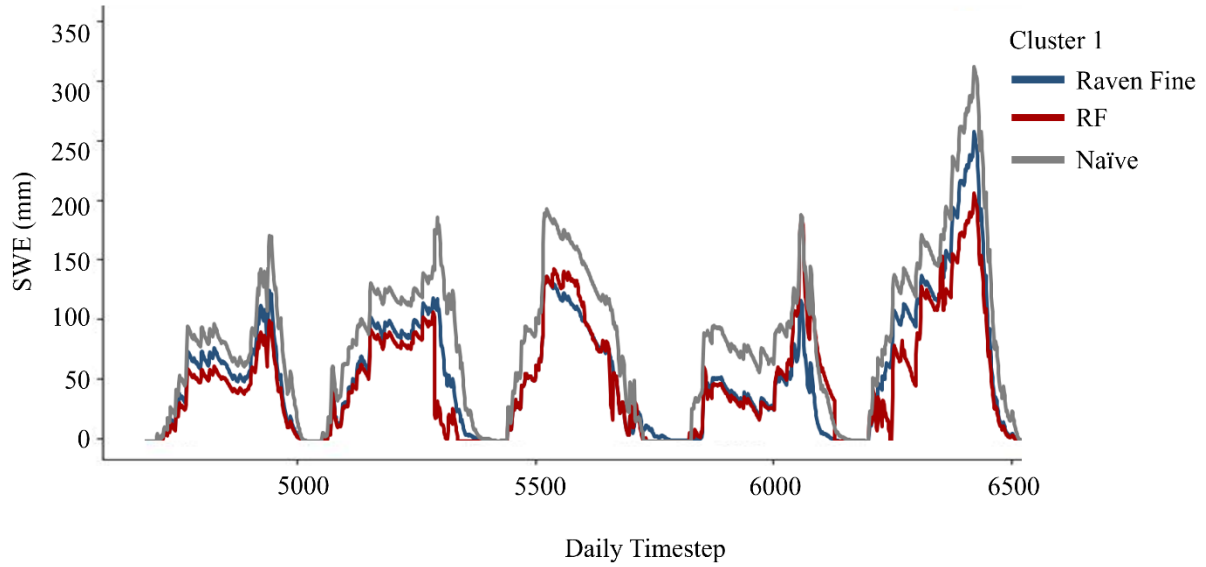


Figure 4.9: Presents a time series comparison of Raven fine-resolution SWE, Oldman mass balance estimated SWE from the RF model, and the naïve approach.

For SCA, the naïve upscaling model shows lower KGE and R^2 values compared to the RF model. Both models exhibit a similar MBE near zero, suggesting that neither upscaling approach consistently overestimates or underestimates SCA values. The time series plot for cluster 7 in Figure 4.10 highlights the more abrupt transitions in SCA produced by the naïve model, while the RF model more accurately preserves the temporal patterns of the fine-resolution Raven output. In the case of snowmelt, the RF model demonstrates significantly higher median KGE (0.86) and R^2 (0.68) compared to the naïve model, reflecting strong predictive accuracy and a strong correlation with observed snowmelt rates, as well as a lower MBE. For sublimation, the RF model's KGE, R^2 , and MBE values display narrow ranges, indicating limited variability in performance across clusters. While the naïve upscaling approach shows greater variability in its error metrics, both methods achieve adequate performance in upscaling sublimation.

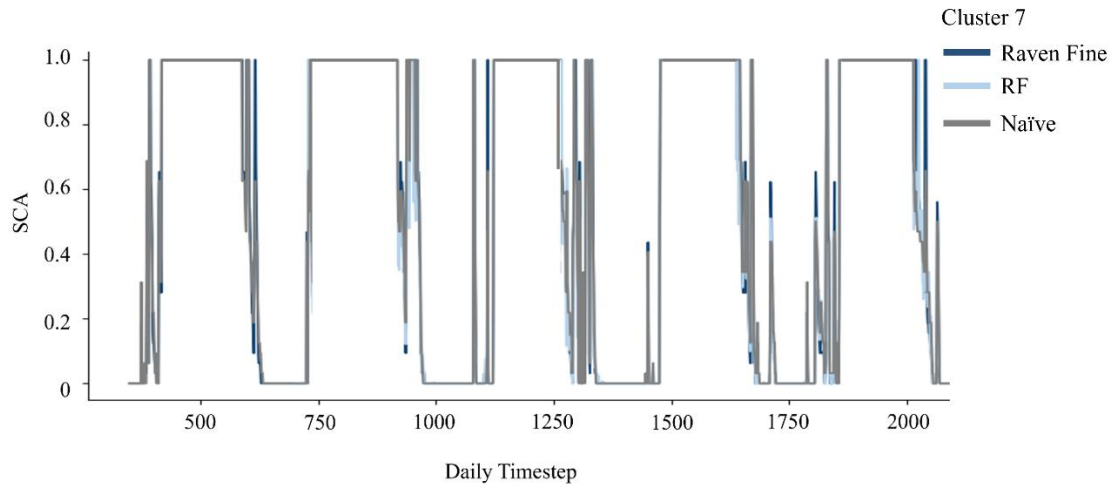


Figure 4.10: Presents a time series comparison of Raven fine-resolution SWE, Oldman mass balance estimated SWE from the RF model, and the naïve approach.

Both the naïve upscaling model and the RF model performed comparably well, with the naïve model showing a tendency to overestimate snow accumulation and ablation processes. Overall, the performance of the RF model suggests that the constitutive relationships derived from the Kananaskis Basin, when applied to the Oldman River Basin, produced acceptable results, providing preliminary support for the transferability of upscaling relationships between areas with similar characteristics. Future research should focus on developing these relationships within a larger basin with more heterogeneous landcover, as the Kananaskis Basin is relatively homogeneous with forested and alpine landscapes. Additionally, deriving relationships specifically for the Oldman Basin and testing their applicability in Kananaskis or another subbasin located within the Bow River Basin in Alberta, Canada could improve our understanding of model transferability across diverse environmental conditions.

Model performance varied notably across clusters, with the RF model achieving its highest KGE values in clusters 10 and 11, which are characterized by mid-to-high elevations and dominant forest and alpine cover. These conditions closely match those in the Kananaskis Basin, where the original upscaling relationships were derived, allowing for stronger generalization. In contrast, clusters 1 and 5 yielded the lowest predictive accuracy. These clusters include agricultural areas, which were absent from Kananaskis, as well as larger proportions of grassland and shrubland, land cover types that were only minimally represented in the original training data used to derive the upscaling relationships. They also contain more low-elevation HRUs, which are associated with warmer temperatures, reduced snowpack accumulation, and greater variability in melt timing. Snow in grassland- and shrub-dominated areas tends to accumulate in thinner layers and melt more rapidly due to minimal canopy cover and higher exposure to radiation. Addressing this variability, future studies could improve the transferability of upscaling relationships by expanding the training dataset to include a broader range of landcover types and elevation conditions.

4.4 Conclusion

This study set out to further investigate the choices made in applying the ML upscaling methodology developed in Chapter 3. examining key components such as the ML algorithm selection and number of domains. Additionally, it explored the feature importance as a proxy for model sensitivity, which may offer deeper insights into key controls on scaling. Through a series of spatial validation experiments, the study rigorously evaluated choices made in the ML upscaling approach under a range of conditions, examining its effectiveness across different algorithms, scales, and variable combinations. In the initial *ML Algorithm Comparison* experiment, the RF model consistently outperformed the XGBoost and ANN models, demonstrating a high level of accuracy and reliability in its predictions while minimizing systematic bias. This superior performance underscored the RF model's ability to capture complex snow dynamics across landscapes with variable topographic and landcover characteristics. Given its robustness in accurately representing snow-related processes, the RF model was selected for all subsequent experiments.

The experiments further explored the conditions under which the RF constitutive relationships are valid. The *Spatial Resolution Comparison* evaluates the constitutive relationships over a range of scales by progressively refining the domain granularity from 2 to 64 domains per cluster. Beyond 16 domains, the increased number of domains led to diminishing returns in predictive performance. This finding suggests that the optimal resolution strikes a balance between preserving data diversity for training. Too few domains can lead to overgeneralization, while too many may reduce variability, hindering effective pattern recognition. Such insights play an essential role in applying the ML upscaling methodology to additional watersheds, particularly those which are more homogeneous in landcover and topography. This result is likely specific to Kananaskis, and when applying the ML upscaling methodology to other watersheds, a similar process should be conducted to determine the ideal resolution for that specific watershed. Future work should look at evaluating how the ML-upscaling methodology performs in watershed more homogeneous in landcover and topographic characteristics to identify potential limitations.

The *Variable Reduction* experiment identified snow temperature, initial SWE, net shortwave radiation, initial SCA, and the distributional information of net shortwave radiation as the key controls on snow accumulation and ablation upscaling relationships in mountainous terrains. The significance of these variables reflects the RF algorithm's ability to capture the physics embedded in the fine-scale snow. However, the RF algorithm can only learn relationships present within the data. If the equations used within the fine-scale model were changed, it is likely that these changes would be mirrored in the variable significance rankings, reflecting the RF algorithm's adaptation to a new representation of physical dynamics. Nonetheless, the identified scaling relationships are consistent with observations from the literature, supporting the model's capability to capture key attributes tied to landcover and terrain.

The adequate performance of the *ML Relationship Transferability* experiment, which applied RF-derived constitutive relationships from the Kananaskis to the Oldman River Basin, highlights the potential for ML-derived constitutive relationships to extend beyond their initial training dataset to areas with similar topographic and landcover characteristics. By capturing essential scaling relationships, the ML upscaling methodology outperforms naïve upscaling, offering a practical solution for simulating snow accumulation and ablation processes in mountainous terrains. Future research should investigate the degree of similarity in topographic and landcover characteristics necessary for the successful transfer of these ML-derived relationships. Understanding this could clarify the limitations of ML-based upscaling in new areas and help establish guidelines on when such relationships can be reliably applied.

Chapter 5

Downscaling snowpack evolution processes with machine learning

5.1 Introduction

The spatial and temporal distribution of snow plays a fundamental role in driving the hydrological cycle and influencing climate in cold regions. Fine-resolution estimates of snow water equivalent (SWE) and snow ablation processes are useful for reducing uncertainties in hydrological models (Raleigh et al., 2015; Vionnet et al., 2022), which in turn can enhance flood management and drought mitigation (Barnett et al., 2005; Dierauer et al., 2018), avalanche hazard forecasting (Morin et al., 2020), and hydropower generation (Lundquist and Dettinger, 2005). To better capture spatial variability in SWE, hyper-resolution snow models, operating at scales of 1 to 100 meters, are increasingly used for water management and design decisions. These models explicitly represent the complex interactions between topography, vegetation, and meteorology that drive snow accumulation and ablation processes, particularly in cold regions (Marsh et al., 2020). However, it is computationally expensive to deploy high-resolution models to try and represent these complex snow processes. To overcome this limitation, computational downscaling may be used to generate fine-resolution estimates from coarser resolution model results.

Downscaling is the process of translating coarse or large-scale data or model outputs into finer, higher-resolution datasets (Najafi et al., 2010). Statistical downscaling is a widely used methodology for generating high-resolution datasets by establishing statistical relationships between coarse-scale and fine-scale variables. For example, Moradizadeh et al. (2023) developed a machine learning-based downscaling

algorithm using Support Vector Regression (SVR) to enhance the spatial resolution of SWE estimates over Iran's Zayandehroud River Basin by integrating passive microwave brightness temperatures, terrain parameters, and vegetation cover. Data assimilation is another common downscaling approach for snow, where fine scale observational data (e.g., remote sensing, in-situ measurements) is used to refine coarser-scale model outputs. Smyth et al. (2020), for example, used a particle filter data assimilation method to downscale snow depth observations across various climates in the western United States. The choice of downscaling technique often depends on data type and availability, modelling objectives, and regional characteristics,

Linear regression was once the dominant method for statistical downscaling of hydrologic variables (Huth, 2002; Zorita and von Storch, 1999). However, due to the complexity and non-linearity of relationships between variables, often driven by system heterogeneity, linear regression can struggle to deliver satisfactory downscaling results (Ghosh and Mujumdar, 2008). Consequently, machine learning (ML) methods have become prevalent in statistical downscaling research. These methods can effectively capture non-linear relationships by extracting patterns from the data, allowing them to generalize these relationships beyond the training dataset (Chaudhary et al., 2022). Among the various ML algorithms, tree-based models have proven highly effective in downscaling applications. Particularly, RF (Breiman, 2001) and XGBoost (Chen and Guestrin, 2016) are popular in the downscaling literature due to the ability to deal with complex nonlinear relationships between variables while minimizing problems with overfitting. For example, Schneider et al. (2022) applied RF models to downscale hydrologic model outputs of groundwater table depths from 500m to 100m resolution, training on simulated climate change impacts from high-resolution hydrological sub models and auxiliary data from the 500m model, validating these outputs against high-resolution hydrological models for sub catchments in Denmark. Whereas Folberth et al. (2019) trained an XGBoost model on low-resolution global crop simulations to estimate maize yield at a finer spatial resolution for regions in Mexico. However, most ML-based downscaling studies rely on observational data rather than model outputs to train their ML models. For instance, Yang et al. (2017) used RF regression to downscale land surface temperature in arid regions using multiple remote sensing indices, and Wunsch et al. (2022) employed deep learning to downscale groundwater storage data in Germany, relying primarily on in situ well observations.

Capturing local heterogeneity is crucial for preserving the spatial variability and complexity of a system, ensuring that key localized factors influencing hydrological processes are accurately represented at both coarse and fine resolutions. Burdett and Craig (in review) developed an ML upscaling methodology that offers a novel approach to estimate spatially averaged hydrological fluxes and states while accounting for system heterogeneity at a resolution smaller than the computational scale. This method involves first running a fine-resolution hydrological simulation to capture detailed states and fluxes, which are then spatially aggregated. An ML model is trained on this aggregated data, using low-order statistical moments of state and forcing variables to retain important distributional information.

The ML model can then be applied to rapidly simulate coarse-resolution mean fluxes over an independent time period, which can be validated against the spatially averaged outputs of the fine-resolution model. A similar ML methodology can be applied to downscaling by reversing the process: instead of aggregating fine-resolution outputs, an ML model can learn to disaggregate coarse-resolution data into finer spatial scales. The model can ensure that the influence of fine scale characteristics, such as variations in topography or landcover, are accurately represented in the downscaling outputs. This study evaluates how effective the physics-informed ML strategy is at downscaling, examining its ability to generate accurate fine-scale estimations of SWE and snow fluxes. Here, analysis is limited to model-generated coarse and fine scale SWE and fluxes: i.e., can we downscale coarse-scale model outputs such that they are consistent with those generated by a fine scale model?

This study looks to develop and test a physics-informed downscaling methodology to derive fine spatial resolution SWE and snow flux estimates from coarse resolution simulated SWE using ML techniques. A case study in the mountainous watershed of the Kananaskis River Basin serves as a foundation to assess the ML downscaling methodology. The objectives of this study are to (1) develop, apply, and test a novel ML downscaling methodology to examine and characterize relationships between, states, fluxes, and across the land surface at different scales, and (2) apply the downscaling methodology to refine the outputs from two distinct upscaling models: a coarse-resolution homogeneous snow model and the ML upscaled snow model of Burdett and Craig (in review).

5.2 Methods

5.2.1 Study Area and Data

The methods developed herein are applied to the Kananaskis River Basin, located in southwest Alberta, Canada, within the Bow River Basin on the eastern slopes of the Rocky Mountains. The Kananaskis River Basin, defined as the region upstream of water survey of Canada gauge 05BF025, spans approximately 955 km² and ranges in elevation from 1271 to 3205m. The climate is characterized by significant variability due to the elevation and topography, generally exhibiting a cool, dry climate with most precipitation occurring in May and June (208 mm), with July and August collectively averaging 130 mm of rain (Whitfield, 2014). Landcover in the Kananaskis River Basin transitions from mixed forests at lower elevations to coniferous subalpine forests and alpine grasslands at higher elevations, with sparse vegetation above the tree line (Evans et al., 2021).

This study uses daily resolution Daymet meteorological data (Thornton et al., 2021) from 1995 to 2020, available at a 1 km resolution, which includes precipitation, shortwave radiation, maximum and

minimum air temperature, and vapor pressure. Wind speed was derived using WindMapper, a wind downscaling method by Marsh et al. (2023), which adjusts 9 km resolution ERA5-Land reanalysis (Herbasch et al., 2020) u and v wind components across varied terrains using the WindNinja (Wagenbrenner et al., 2016) model to account for topographic effects to derive downscaled windspeed at a resolution of 1 km. Within the Raven hydrological model (Craig et al., 2020), a simple linear elevation lapse approach is applied to correct Daymet precipitation and temperature to the HRU scale. This method adjusts temperature and precipitation values based on the HRU's elevation relative to the mean elevation of the forcing grid cell, ensuring localized accuracy in response to elevation differences. The elevation data was obtained at a 25 m resolution using the R `elevatr` package (Hollister et al., 2023), which leverages Shuttle Radar Topography Mission (SRTM) data (U.S. Geological Survey, 2014). Slope and aspect were also derived from this elevation data. Additionally, landcover data was acquired from Natural Resources Canada (NRCan) (2015) and used for the delineation of Hydrologic Response Units (HRUs). Delineation of the Kananaskis River basin was carried out using BasinMaker (Han et al., 2023) within the Magpie workflow (Burdett and Craig, 2024), an open-source tool for generating input files for the Raven hydrologic model. BasinMaker used the hydrometric flow gauge 05BF025 within the North American Lake-River Routing Product (Han et al., 2021) to define the study area drainage extent. The digital elevation model (DEM) data is then used to identify topographic features, build a flow network, and delineate subbasin boundaries. The initial HRU discretization was generated based on elevation, aspect, and landcover, and a grid overlay created approximately 27,000 finer-resolution HRUs, nearly uniform in area (each $\sim 0.035 \text{ km}^2$).

5.2.2 Machine Learning Downscaling Methodology

Here we propose a novel approach to spatial downscaling that integrates physics with ML methods, enabling the direct downscaling of both state and flux variables while preserving mass balance. The downscaling methodology proposed and tested here includes the following steps, as shown in Figure 5.1: (1) a fine-resolution hydrological model (the “training model”) is used to simulate detailed states and fluxes over a subset of HRUs in space and time that are then spatially aggregated to a coarser spatial resolution; (2) scaling ratios are then computed as the ratio of the mean states or fluxes at the fine scale to the mean state/flux at the coarse scale for each corresponding time step; (3) an ML algorithm is trained with the fine-resolution topographic and coarse-resolution forcing variables to estimate the scaling ratios for any HRU in any time step; (4) the ML downscaling ratios are then multiplied by the coarse-scale state/flux data to generate the fine-resolution estimates in domains not used for training. These fine-resolution results are then compared to independent results from the fine-resolution hydrological model for validation. The goal is for the local ML-derived ratios to effectively transfer the detailed spatial characteristics from fine-scale estimates to a coarse-scale, enhancing the effective resolution of the coarse-

scale dataset while removing the need for high-resolution simulation at each time step or spatial unit. Notably, lateral exchange between HRUs (e.g., from blowing or avalanching snow) is not simulated in the fine-scale model, and therefore its influence will also not be captured by the ML model.

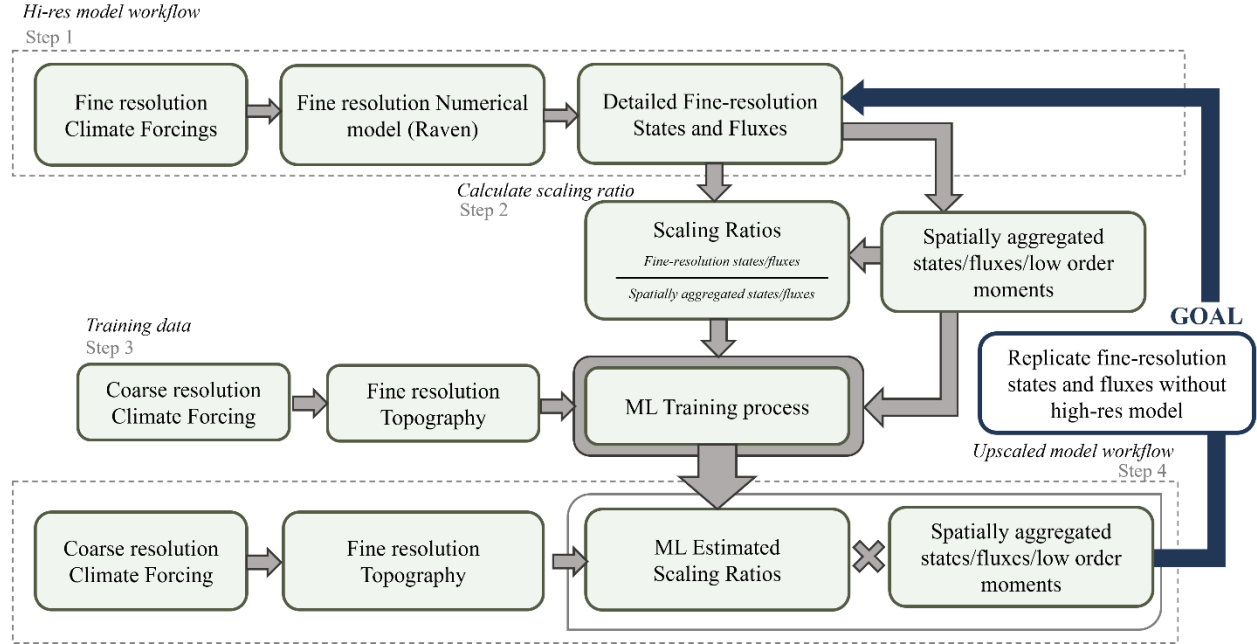


Figure 5.1: Conceptual outline of the steps used in the machine-learning based downscaling methodology.

5.2.3 Fine-Resolution Model

This case study focuses on using the ML downscaling methodology to effectively simulate the fine scale interactions between the atmosphere and local terrain features that influence snow-related state variables such as SWE and SCA. In-situ measurements of these state variables, as well as other snow-related forcing variables like snow temperature, are unavailable at the spatial and temporal resolution required to train ML models. As such, the first step of the ML downscaling methodology involved using a fine-resolution Raven hydrological model to simulate the required variables in a subset of the time and scale domain. A key assumption is that this model is appropriate for representing fine-scale fluxes. The Raven model employed in this study is the same fine-resolution model as described in Burdett and Craig (in review), where it was thoroughly validated and integrated into an ML upscaling methodology framework. This Raven model incorporates a well-tested two-layer snow balance model (Wigmosta et al., 1994) to simulate snow ablation through melt and sublimation, with the energy required for snow melt estimated using an energy balance approach ignoring blowing snow and density changes (Dingman, 2015). Notably, the methodology could readily use alternate fine-scale snow models; the primary

requirement is that it be appropriately sensitive to fine scale forcing inputs, as would be expected of any snow model.

5.2.4 Scaling Ratios

The scaling ratio characterizes the relationship between local values and their corresponding coarse-scale averages generated using a larger resolution model. For each fine-scale grid cell, the scaling ratio (δ_i) is computed as the ratio of the fine-scale SWE (SWE_i) to the coarse-scale SWE (SWE_A). A ML model is trained to then predict these scaling ratios for each fine-scale grid cell in each time step, allowing for calculation of SWE at the fine scale. Similarly, scaling ratios were also generated snowmelt and sublimation independently and applied to their respective fluxes.

A constraint is applied to the scaling ratios in the downscaling process to maintain consistency with mean coarse scale inputs. The weighted sum is expressed as follows:

$$\frac{1}{N} \sum_{i=1}^N \delta_i = 1 \quad (5.1)$$

where i represents each fine-scale grid cell, N is the total number of cells within the coarse scale region, δ_i is the scaling ratio for the i^{th} cell, and the equation ensures that the sum of all scaling ratios equals one. This constraint assists in preserving the total mass or mean value of the local variable when transitioning from coarse to fine scales. By ensuring that the weighted sum equals one, the methodology maintains the overall mass balance, preventing any artificial increase or decrease in the total SWE during the downscaling process while also maintaining consistency between coarse- and fine-scale solutions. This constraint was implemented in XGBoost by incorporating a custom objective function that penalizes any deviation from a weighted sum of one, guiding the model to satisfy this condition while optimizing predictions.

5.2.5 Machine Learning Algorithm

In the third step of the ML downscaling methodology, the ML algorithm is trained using fine-resolution topographic variables (elevation, slope, and aspect) along with coarse-resolution states and fluxes to estimate the scaling ratios. The ML-estimated ratios are then multiplied by the coarse-resolution states and fluxes to downscale them to the fine resolution. The downscaled model results are then validated against the fine-resolution model values from an independent dataset to assess the ML algorithm's ability to capture the spatial variability of state and flux variables at fine scales. This validation

is applied both in space and time. This ensures that the downscaling method effectively bridges the gap between coarse-resolution inputs and fine-resolution observations.

XGBoost is an optimized gradient boosting algorithm which builds an ensemble of decision trees sequentially, leveraging residuals from previous trees to progressively enhance model accuracy (Chen and Guestrin, 2016). XGBoost was selected as the ML method for this study due to its previous use in downscaling applications (Pang et al., 2017) and its ability to handle complex datasets, capture nonlinearity, and provide robust performance despite the presence of outliers, noise, and correlated variables. Additionally, XGBoost allows for the implementation of custom constraints, such as the weighted sum constraint, within its objective loss function. This flexibility enables the model to enforce conditions like maintaining the sum of scaling ratios to equal one, a process critical for mass conservation in downscaling. RF models lack this ability, as it constructs each decision tree independently, without an optimization process that allows for integrating custom constraints. For this study, we employed the Python Scikit-learn XGBoost regression algorithm. The XGBoost model underwent a training and validation (1994-2000) period during which hyperparameters were manually adjusted to produce the final architecture of the XGBoost model. The final XGBoost model consisted of 125 decision trees, each pruned to a maximum depth of 20, and learning rate of 0.05 samples per leaf node.

The XGBoost model is trained on the fine-resolution and coarsely aggregated variables outputs of the hydrologic model, meaning it is effectively learning from physically-based simulations. As a result, the model is informed by the physics embedded within the high-resolution model rather than relying solely on empirical correlations from observed data. Hence, the selection of input variables plays a vital role in the success of the XGBoost model; the omission of a significant variable will likely result in the XGBoost model underperforming. The input and output variables used to train the RF model are indicated in Figure 5.2. The fine-resolution training dataset incorporates elevation, slope, aspect, and incoming shortwave radiation, which are critical variables for accurately capturing spatial variations in SWE. While shortwave radiation is a forcing variable, it can be derived from topographic information such as elevation, slope, and aspect, as these factors influence solar angles and shading effects. This makes shortwave radiation a practical and viable choice to include at the fine resolution, enabling the model to account for localized energy inputs. The remaining forcing and state variables, such as temperature, precipitation, and SWE, are included in the training dataset at a coarse resolution, providing the necessary context to estimate fine-resolution scaling ratios for SWE. The downscaled fine-resolution SWE values are then used to compute snow covered area (SCA) for each elevation band within a domain. SCA is determined by calculating the ratio of the area where SWE is present (exceeding a threshold of 0.25 mm) to the total area of HRUs within the subregion of interest.

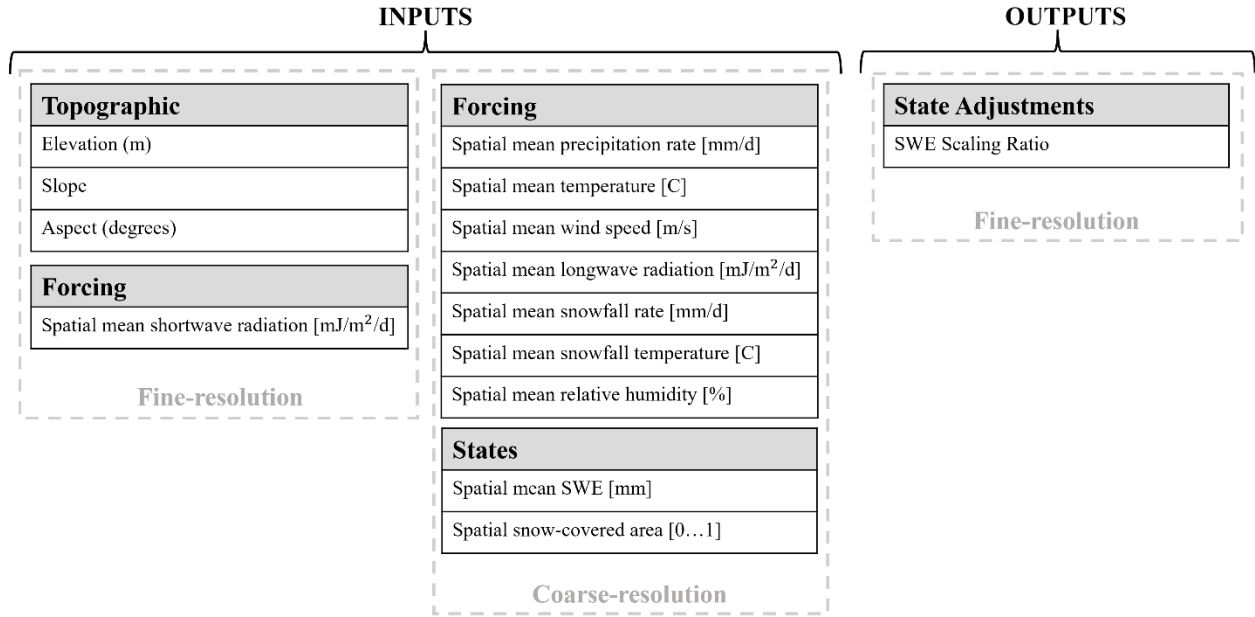


Figure 5.2: The input and output vectors for the XGBoost model. All mean forcing represents spatial averages over the coarse scale.

Four numerical experiments were designed to evaluate the accuracy of the ML downscaling methodology and sensitivity to methodological choices. These experiments include the (1) *Spatial Validation (Ideal)*, (2) *Temporal Validation (Ideal)*, (3) *Spatial Validation (Upscaling Input)*, and (4) *Spatial Validation (Naïve Input)*. The *Spatial Validation (Ideal)* experiment (experiment 1) serves as a spatial validation to evaluate the efficacy of the ML downscaling approach under simplest conditions. This experiment tests whether the downscaling methodology can accurately reconstruct high-resolution model output from spatial mean states at the coarse scale. A k-means clustering algorithm is used to define 12 clusters that share similar topographic and landcover characteristics (MacQueen, 1967); these are the same clusters used in Chapter 4. Figure 5.3 illustrates the division of the Kananaskis basin into 12 clusters. Each of these clusters is further divided into 5 equal-area coarse-scale contiguous domains at random for testing. The HRUs from two of these domains are used to train the XGBoost algorithm on 14 years of data (2000-2014), to then estimate the scaling ratios for the three remaining testing domains. After estimating these ratios with the XGBoost model, the domain scale SWE is multiplied by the scaling ratios to downscale them to the HRU resolution and compute the downscaled SCA. The performance of this downscaling methodology is then assessed by comparing the downscaled SWE, SCA, snowmelt, and sublimation variables to the original fine-resolution estimates from the Raven model for each HRU. The downscaled SCA values, matching the same spatial extent as the computed downscaled SWE values, are then compared to the SCA derived from the original fine-resolution SWE estimates generated by the Raven model.

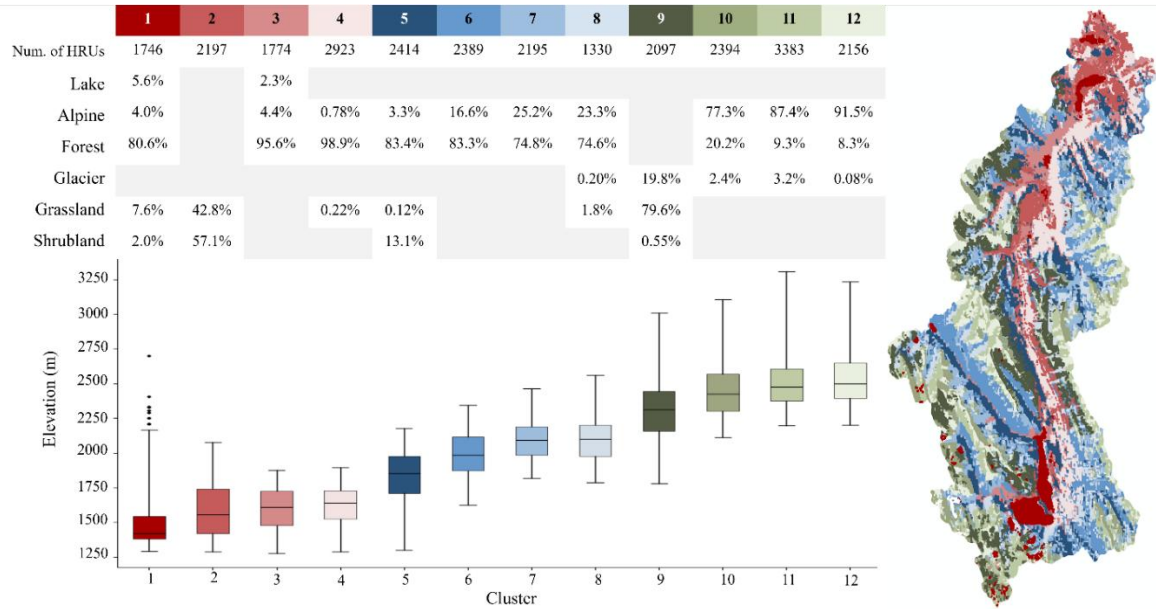


Figure 5.3: Visualization of the 12 k-means clusters in the Kananaskis River Basin, which are the same clusters used in Chapter 4 accompanied by a table detailing the number of HRUs per cluster, elevation distribution, and landcover percentages.

The *Temporal Validation (Ideal)* experiment (experiment 2) is conducted to enhance the robustness of the downscaling methodology by evaluating its predictive accuracy over an extended period. In this experiment, the XGBoost model is trained on 14 years (2000-2014) of data across all five domains. The model includes shortwave radiation and topographic variables at the HRU scale, as well as state variables at the domain scale, to estimate scaling ratios for SWE for the remaining 6 years (2014-2020). The 70/30 temporal training split ensures that the model has adequate data diversity within the training dataset for effective learning while retaining a sufficient portion of the dataset for evaluation. This approach evaluated whether the XGBoost model successfully captures temporal variations in the scaling ratios. The downscaled SWE and SCA outputs are validated by comparing them to the original Raven fine-resolution model, demonstrating its ability to downscale data on an independent time period.

The scaling ratios derived in the *Spatial Validation (Ideal)* experiment are then used in the *Spatial Validation (Upscaling Input)* experiment (experiment 3) to assess the effectiveness of the ML downscaling methodology when applied to a coarse-scale model output (rather than a spatial average of fine outputs). First, the ML upscaling methodology of Chapter 3 employed a mass balance equation to calculate mean SWE and SCA based on the upscaled snowmelt and sublimation values. These upscaled outputs were then downscaled by applying the corresponding scaling ratios trained in the *Spatial Validation (Ideal)* experiment, reconstructing SWE back to HRU resolution. The final downscaled outputs were compared to the fine-resolution Raven model outputs for each HRU.

The second upscaling methodology estimated coarse-scale state and fluxes by applying point-scale relationships to subbasin average forcings and parameters in the *Spatial Validation (Naïve Input)* experiment (experiment 4). This naïve upscaling method assumed homogeneity, ignoring complex interactions and heterogeneity across scales. Here, each cluster's domains were treated as single equivalent HRUs in a simplified Raven model to estimate mean SWE and SCA. The ML downscaling ratios from the *Spatial Validation (Ideal)* experiment were applied to the naïvely upscaled snow model estimates to reconstruct the values back to the HRU scale. The reconstructed HRUs were evaluated against the original fine-resolution Raven estimates.

The error metrics selected for this study include the Kling-Gupta Efficiency (KGE), Mean Absolute Error (MAE), and coefficient of determination (R^2), applied to pairwise differences in all fine resolution HRUs at all times. The metrics chosen can be used together to simultaneously describe the overestimation/underestimation and magnitude of the error. KGE is typically used to assess the overall agreement between observed and predicted values in a time series, taking into account the correlation, bias, and variability (Gupta et al., 2009). KGE values are calculated for the entire time series within each HRU, rather than as a mean of individual HRU values. The result is a single KGE value for each HRU that reflects the overall agreement. MAE quantitatively reflects the degree of dispersion of the predictions by calculating the average of the absolute errors between predicted and actual values (Chai and Draxler, 2014). MAE was selected instead of MBE because the imposed mass balance constraint prevents any artificial increase or decrease in total SWE during the downscaling process, ensuring consistency between coarse- and fine-scale estimates. As a result, MBE which quantifies average over- or under-prediction, becomes less informative since the global bias is effectively forced to zero. In contrast, MAE captures the average magnitude of local errors, making it more suitable for evaluating how well the downscaled values match the fine-scale distribution while still respecting the overall mass balance. R^2 evaluates the proportion of the variance in the observed data that is explained by the model, providing a measure of how well the model captures the variability in the data. Low values of MAE suggest more accurate predictions, while higher KGE and R^2 values (closer to 1) indicate better overall model performance.

5.3 Results and Discussion

5.3.1 Spatial Validation

In the *Spatial Validation (Ideal)* experiment (experiment 1), the XGBoost model trained on one set of spatial domains is used to estimate the scaling ratios to reconstruct the fine-resolution outputs in other domains within the same cluster from coarsely aggregated input data. This experiment is designed

to assess the ability of the ML downscaling methodology to preserve fine-scale variability after variables have been coarsely aggregated. For each of the 12 clusters, the XGBoost algorithm is trained on three of the five domains using fine-resolution topographic variables and shortwave radiation alongside forcing and state variables that have been spatially aggregated to the domain scale. Each cluster is trained independently, yielding a unique ML model for each cluster. The ML model then estimates scaling ratios at the fine spatial resolution for the remaining two domains HRUs across a 14-year study period. The coarsely aggregated state variables at the domain scale are subsequently multiplied by the ML-estimated scaling ratios to downscale these variables back to the fine-resolution HRU scale. This process provides a direct evaluation of the downscaling methodology using ‘ideal’ coarse-scale inputs, in contrast to other experiments in this study that examine naively and upscaled coarse-scale inputs, allowing for an assessment of the model’s efficacy in reconstructing detailed spatial variability.

5.3.1.1 Downscaling SWE

All 12 clusters achieved median KGE values above 0.60, indicating a moderate to high level of agreement between the original fine-resolution Raven SWE and the XGBoost-reconstructed SWE, as shown in Figure 5.4. The narrow interquartile ranges further highlight the consistency of KGE values across clusters. Median R^2 values exceeded 0.80 for all clusters, suggesting the model effectively captured the variance in the original SWE data. Additionally, median MAE values remained below 10 mm across all clusters, demonstrating consistently low absolute errors between the reconstructed and fine-resolution Raven SWE. These results suggest that the *Spatial Validation (Ideal)* experiment effectively preserved the spatial heterogeneity of SWE during downscaling.

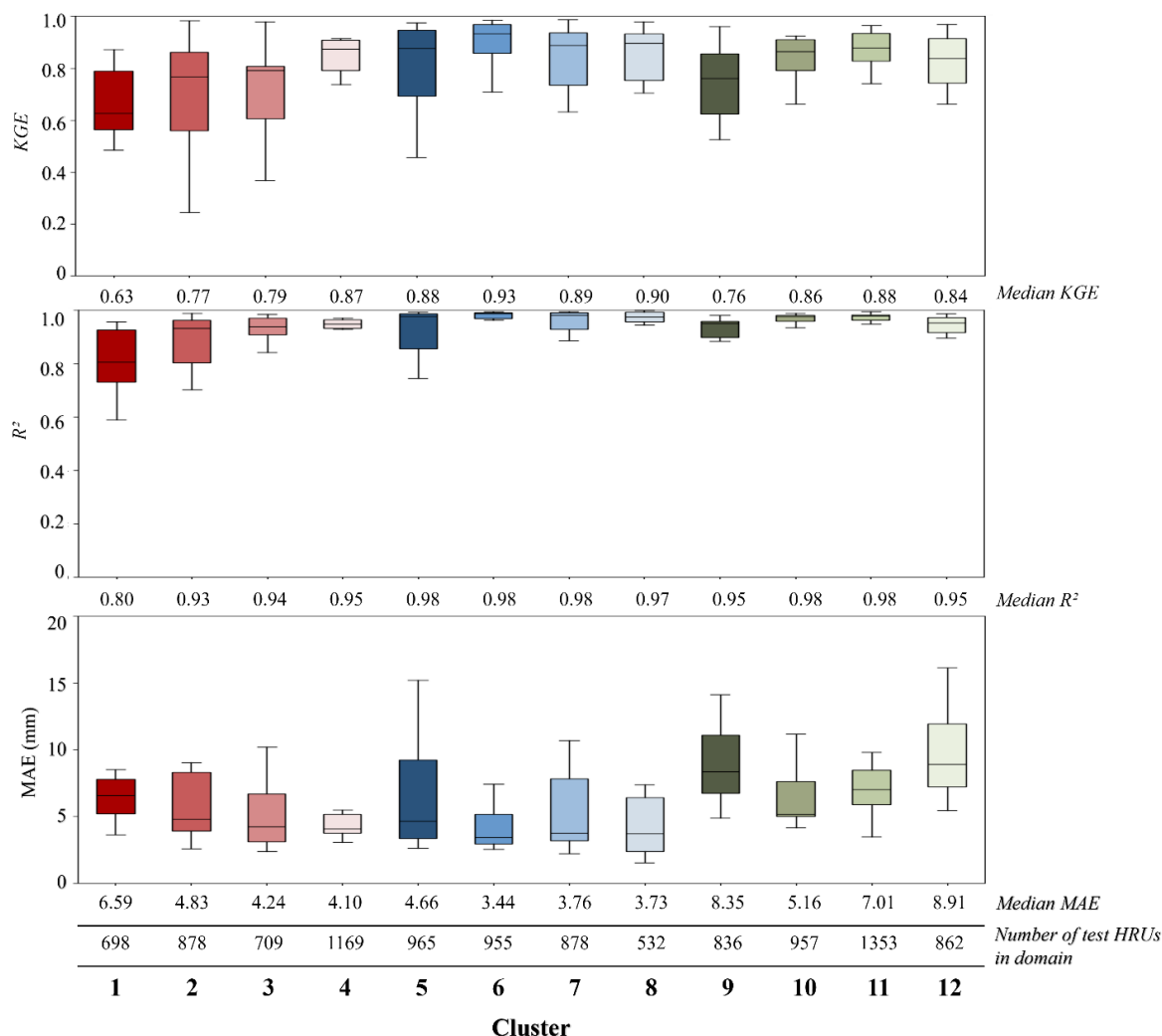


Figure 5.4: Illustrates the KGE, R^2 , and MAE values for the 12 clusters, assessing the agreement between the ML-downscaled SWE values and the original Raven fine-resolution estimates in the Spatial Validation (Ideal) experiment.

However, cluster 1 exhibits a wider range of KGE and R^2 values, including some of the lowest KGE scores among all clusters, indicating greater variability in agreement between the reconstructed and fine-resolution SWE data. Similarly, clusters 9, 10, 11, and 12 show greater variability in MAE values, with consistently higher MAE compared to other clusters. This suggests that while the XGBoost model effectively captures overall SWE patterns, it struggles to accurately reproduce SWE magnitudes within these specific clusters, often exhibiting a tendency to either overestimate or underestimate values. For example, Figure 5.5 presents a line plot of three downscaled HRUs from a testing domain within clusters 3, 5, and 12. Although this represents only a small subset of the total HRUs in the domains, these plots highlight the variability in model performance during downscaling, showcasing differences across low-, mid-, and high-performing HRUs. In the low and mid performing HRU, the XGBoost model tends to

consistently underestimate the ML-downscaled SWE. In contrast, the high-performing HRU underestimated peak SWE values, although it still closely aligned with original fine-resolution SWE values.

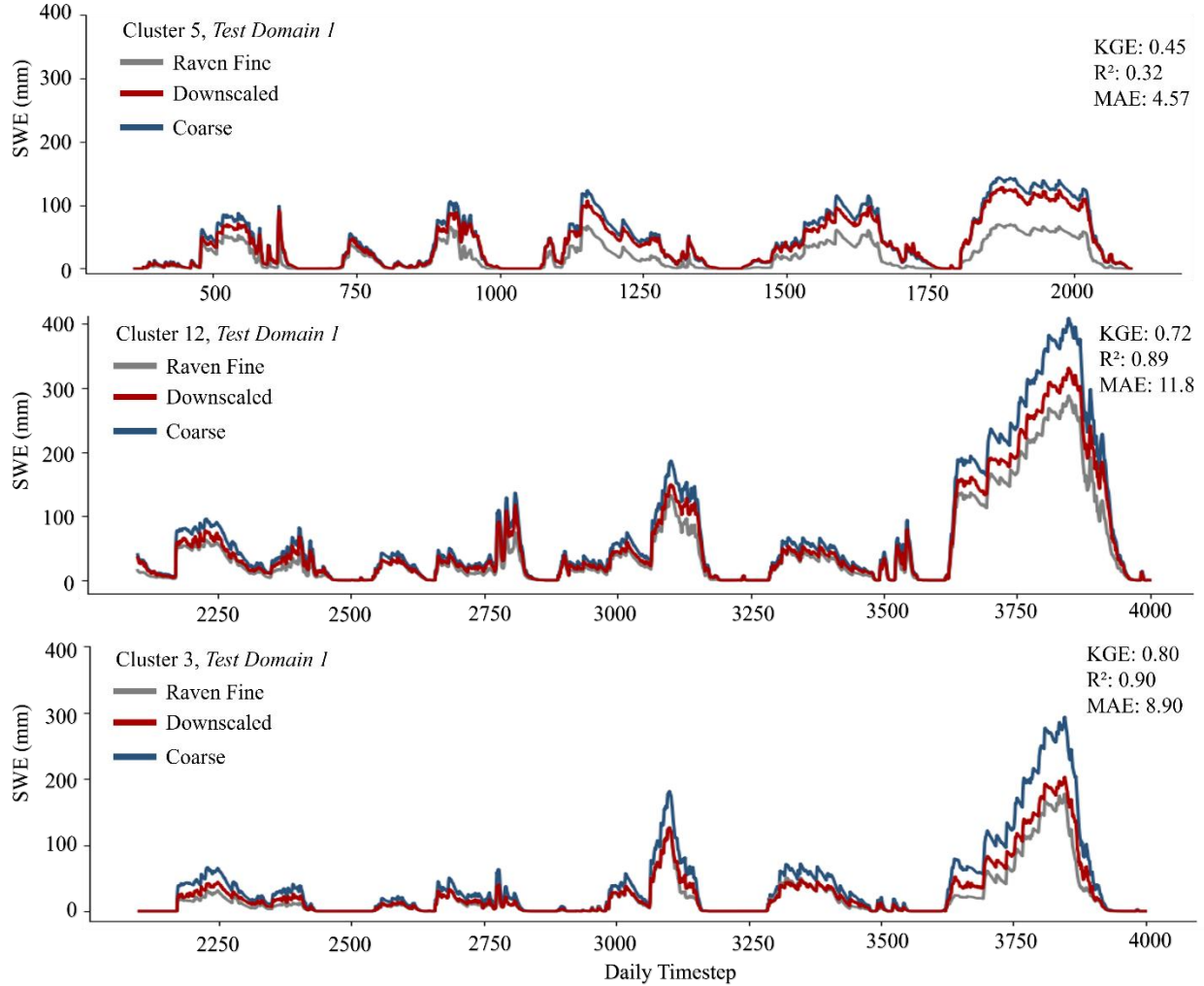


Figure 5.5: Time series plot comparing fine-resolution Raven SWE, coarse-resolution SWE (before downscaling), and ML-downscaled SWE for three HRUs in clusters 3, 5, and 12. The plot illustrates overall agreement, highlighting differences between a low, moderate, and high performing HRU and a high-performing HRU.

Clusters 9, 10, 11, and 12, located at the highest elevations within the study area, exhibit more complex SWE dynamics due to greater SWE accumulation, resulting in higher variability in MAE. This complexity makes these clusters more challenging for the XGBoost model to reconstruct SWE accurately. At higher elevations, snow accumulation tends to be greater due to colder temperatures which promote snowfall instead of rain, orographic lifting that enhances precipitation as moist air rises and cools, and prolonged snow retention due to slower melt rates (Cartwright et al., 2020). In these clusters, south-

facing slopes receive more solar radiation, accelerating melt, while north-facing slopes retain snow for longer periods, creating sharp spatial contrasts that are more difficult for the downscaling model to resolve. Additionally, forested areas moderate snow processes by intercepting snowfall and reducing ablation rates through canopy shading (Varhola et al., 2010). In contrast, open alpine areas are more exposed to wind and solar radiation, leading to enhanced sublimation and snowmelt (Andreadis & Lettenmaier, 2006). These differences in landcover contribute to greater spatial variability in SWE, particularly in non-forested areas.

5.3.1.2 Snow-Covered Area

The ML-downscaled SWE estimates were used to derive the SCA for each domain within a cluster. The ML SCA estimates were compared against the SCA derived from Raven fine-resolution outputs, yielding median KGE values greater than 0.80 across testing domains and clusters, as shown in Figure 5.6. The consistently high median R^2 values further indicate a strong agreement while the predominately low MAE values suggest minimal error between the SCA estimates. However, cluster 1 exhibits higher MAE and lower KGE values, suggesting that the ML model struggles more to capture SCA in this cluster compared to others.

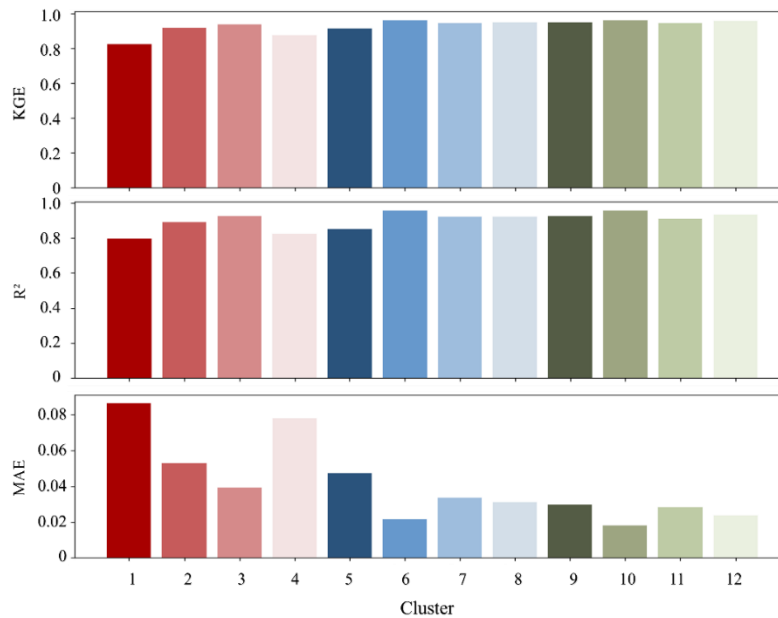


Figure 5.6: Illustrates the KGE, R^2 , and MAE values for one of the testing domains within each of the 12 clusters, assessing the agreement between the ML derived SCA values and the SCA estimated from fine-resolution Raven SWE estimates.

To further evaluate the SCA estimates, the fine-resolution Raven model and ML-derived SCA values were compared against the ERA5-Land SCA dataset. ERA5-Land provides global reanalysis data at

a coarser 9 km resolution, offering daily estimates of surface conditions, including SCA, and serving as a widely recognized benchmark for hydrologic and climate studies (Muñoz-Sabater et al., 2021). Figure 5.7 presents the SCA estimates from 2001 to 2005 for clusters 1 and 3, comparing the fine-resolution Raven model, ML-downscaled, and ERA5-Land SCA. The Raven fine-resolution SCA tends to decrease gradually during melt periods, while the coarser ERA5-Land and ML-downscaled SCA exhibit more abrupt shifts from snow to no snow. Although ERA5-Land was not used to directly evaluate the Raven model, it serves as an independent, observationally constrained benchmark for assessing how well the ML-downscaled products captures large-scale patterns and transitions in SCA. This comparison strengthens the evaluation by showing that the ML model generalizes reasonably well when assessed against an independent dataset. Overall, the ML-downscaled SCA model demonstrated moderate success in reconstructing SCA spatial patterns comparable to an independent reanalysis product, highlighting its ability to effectively characterize snow distribution as well as the timing of snow accumulation and ablation periods.

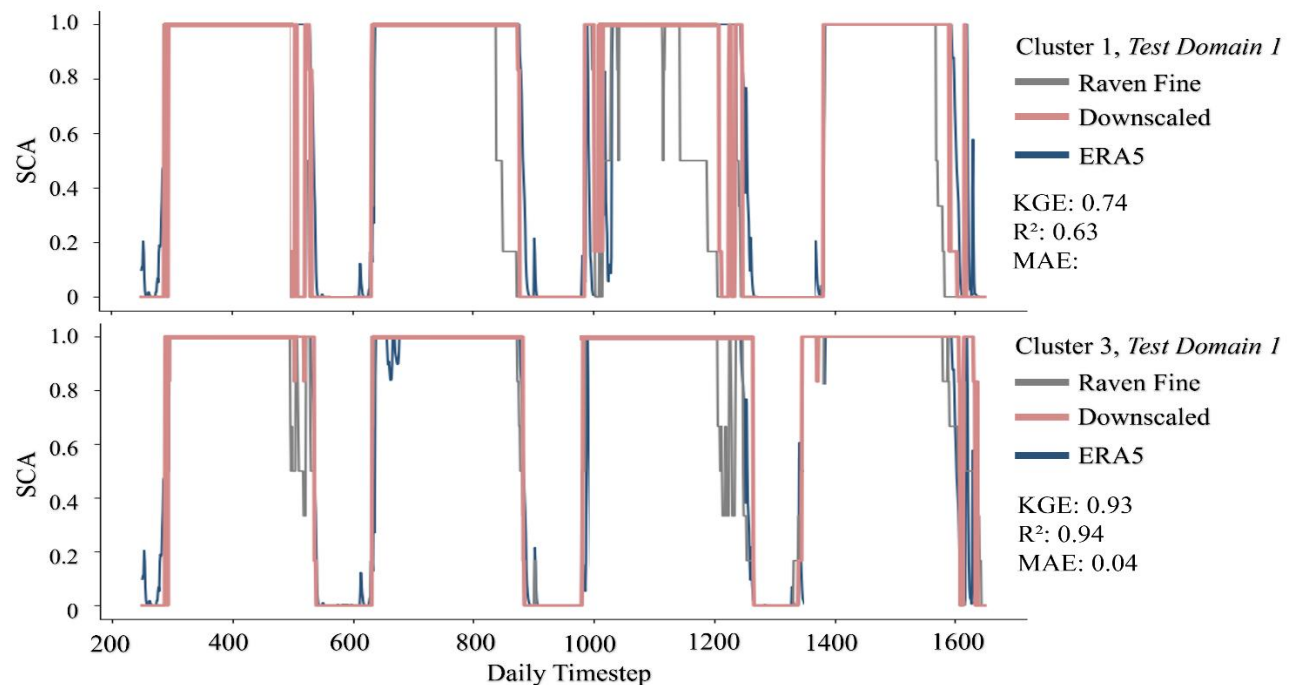


Figure 5.7: Compares the 2001-2005 total SCA estimates from the fine-resolution Raven model, ML-downscaled model, and ERA5-Land dataset.

Accurately capturing the fine-scale snow distribution is valuable for modelling of water availability, especially in regions where snowpack dynamics drive hydrologic processes. The heterogeneity in snow distribution, shaped by factors such as wind exposure, topographic shading, and slope, plays a key role in understanding localized patterns of snow accumulation and ablation. For example, in mountainous regions, wind-sheltered areas tend to retain deep snowpacks, while exposed areas

experience rapid ablation. These localized variations significantly influence downstream water availability (Draeger et al., 2023), highlighting the importance of accurately capturing the distribution of snow when downscaling to ensure reliable estimates of water resources and hydrologic processes. The results from the *Spatial Validation (Ideal)* experiment demonstrates its ability to capture small-scale heterogeneity in SWE and represent localized variations in snow distributions.

5.3.1.3 Extending the ML Downscaling Methodology to Snow-Related Fluxes

The ML downscaling methodology was evaluated using SWE and SCA, though the framework is inherently flexible and has the potential to be applied to other processes, as it is not specific to a single variable. This adaptability was briefly assessed by extending the *Spatial Validation (Ideal)* experiment to downscale additional snow-related fluxes, such as snowmelt and sublimation. The KGE values were again calculated by comparing the ML downscaled snowmelt and sublimation values against their corresponding fine-resolution Raven model outputs. The KGE values showed mixed performance between the two fluxes, with local snowmelt rates achieving median KGE values of approximately 0.80 across most clusters. However, cluster 1 exhibited a noticeably lower median KGE of 0.6, highlighting the model's reduced accuracy in this region, as shown in Figure 5.8. Sublimation, on the other hand, exhibited greater variability in KGE values across clusters, indicating that the model's performance varied depending on regional characteristics. However, the variability in KGE values within individual clusters remained relatively low, as indicated by the narrower interquartile ranges in the box plots in Figure 5.8. This suggests that while sublimation predictions exhibit greater variability across different clusters, the model maintains consistent performance within each cluster.

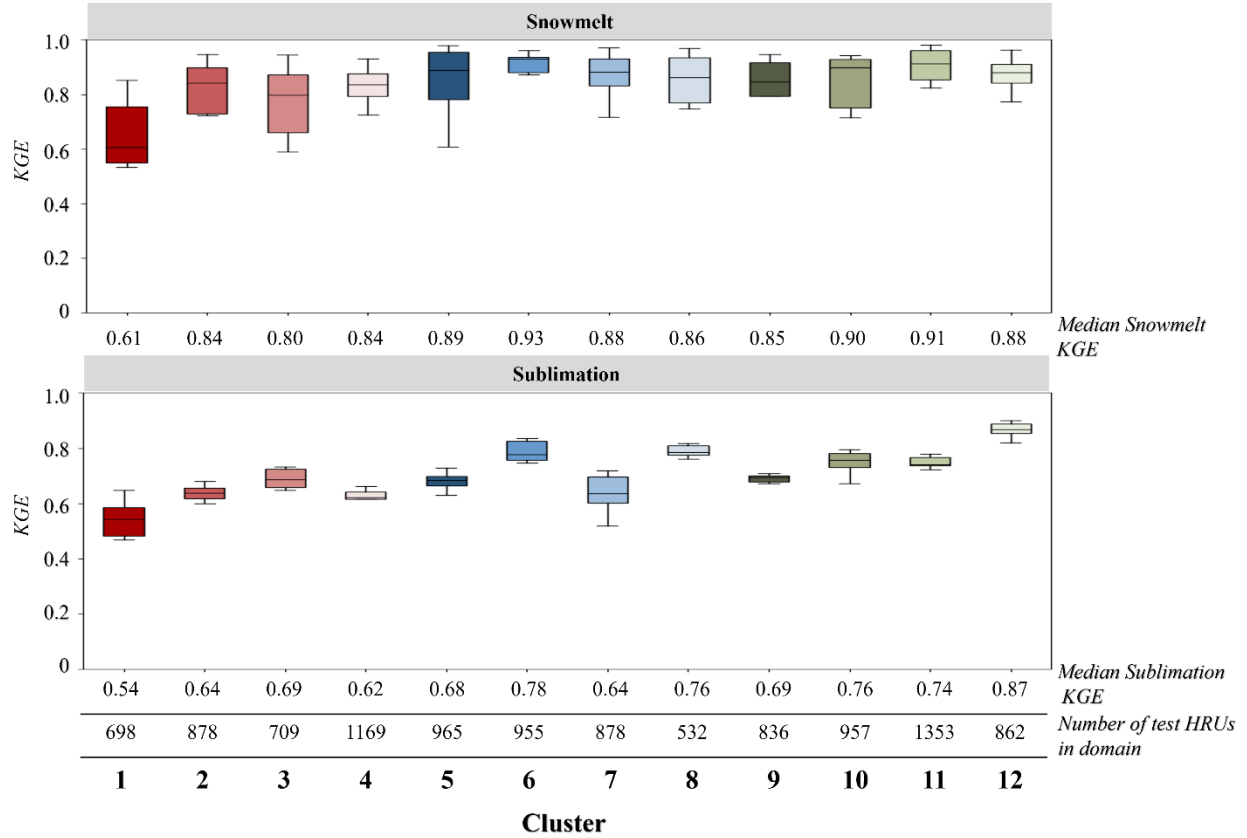


Figure 5.8: KGE values comparing ML-downscaled snowmelt and sublimation fluxes to the original Raven fine-resolution estimates across all 12 clusters.

The ML model demonstrated strong skill in downscaling snowmelt, as illustrated in Figure 5.9, which presents a time series plot of three downscaled HRUs from a testing domain within clusters 3, 7, and 9. These clusters represent low, mid, and high-performing HRUs, providing insight into the model's performance across varying conditions. The XGBoost model effectively captured the overall distribution and trends of snowmelt but struggled to account for the full magnitude of the fine-resolution snowmelt values. For example, cluster 9 ($KGE = 0.59$), underestimated snowmelt, in some instances failing to fully reproduce peak melt events. Despite this, the general timing and occurrence of snowmelt events were reasonably well aligned with the fine-resolution snowmelt. In contrast, cluster 7 ($KGE = 0.92$) exhibited a stronger agreement, where the downscaled SWE closely followed the fine-resolution data, suggesting that in certain clusters, the ML model performed well in preserving both the magnitude and timing of snowmelt events.

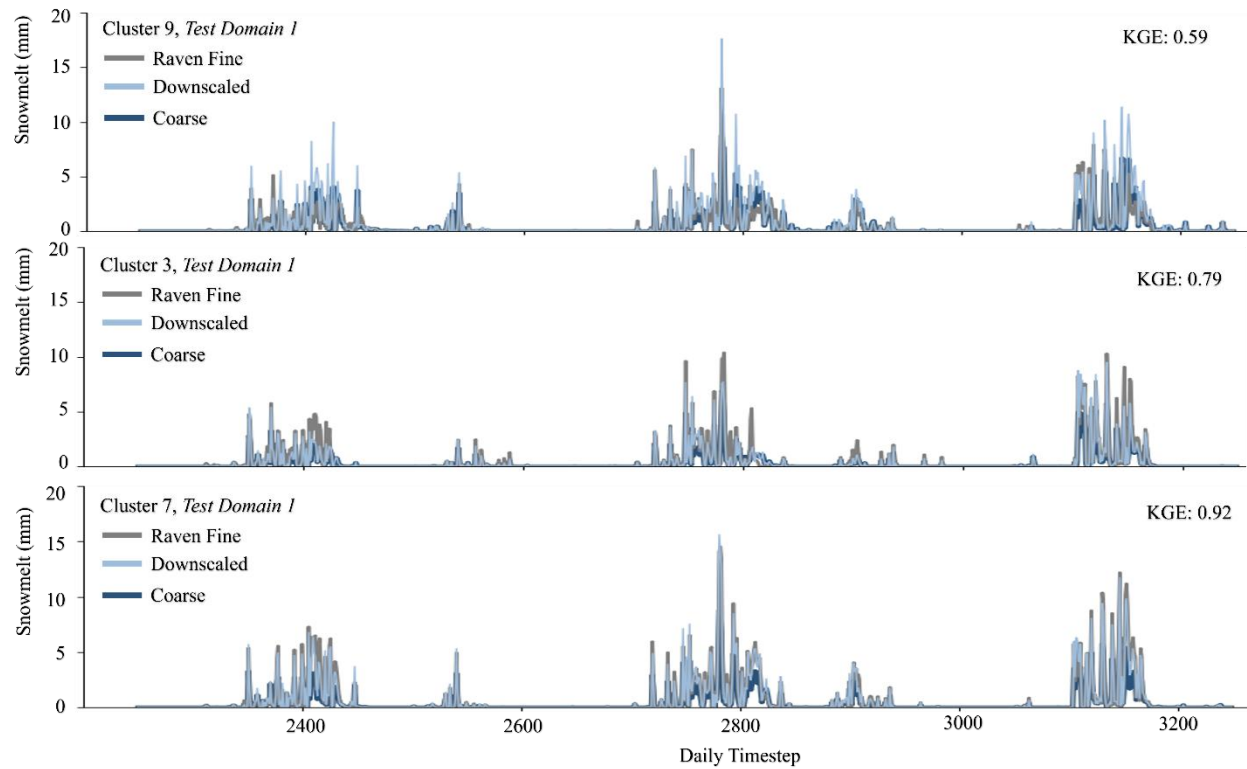


Figure 5.9: Time series plot comparing fine-resolution Raven snowmelt, coarse-resolution snowmelt (before downscaling), and ML-downscaled snowmelt for three HRUs in clusters 3, 7, and 9.

The ML downscaling of sublimation generally exhibited lower KGE values, particularly in clusters at lower elevations. In these clusters, the model frequently overestimated sublimation while also introducing additional noise into the downscaled values, as seen in Figure 5.10 for cluster 1 ($KGE = 0.54$). Whereas cluster 5 ($KGE = 0.62$) achieved a moderate level of agreement between the fine-resolution sublimation and the reconstructed values. However, the downscaled sublimation still exhibits additional noise and a consistent tendency to overestimate sublimation. While the ML downscaling performance improved in higher-elevation clusters as shown by cluster 11, downscaling sublimation is challenged by its reliance on highly variable localized factors, such as wind speed, net shortwave radiation, air temperature, and vapor pressure deficit. These variables exhibit complex spatial and temporal variability, making it harder to model accurately than simpler snow-related processes, like SWE and snowmelt (Reveillet et al., 2020). The ML downscaling methodology faced challenges in accurately reconstructing sublimation, underscoring its limitations in capturing the complexity of snow-atmosphere fluxes, particularly in regions with highly variable environmental conditions. While its success in downscaling snowmelt highlights the broad applicability of the approach, its poor performance in reconstructing sublimation

showcases its success may be dependent on the physical complexity and variability of the process being downscaled.

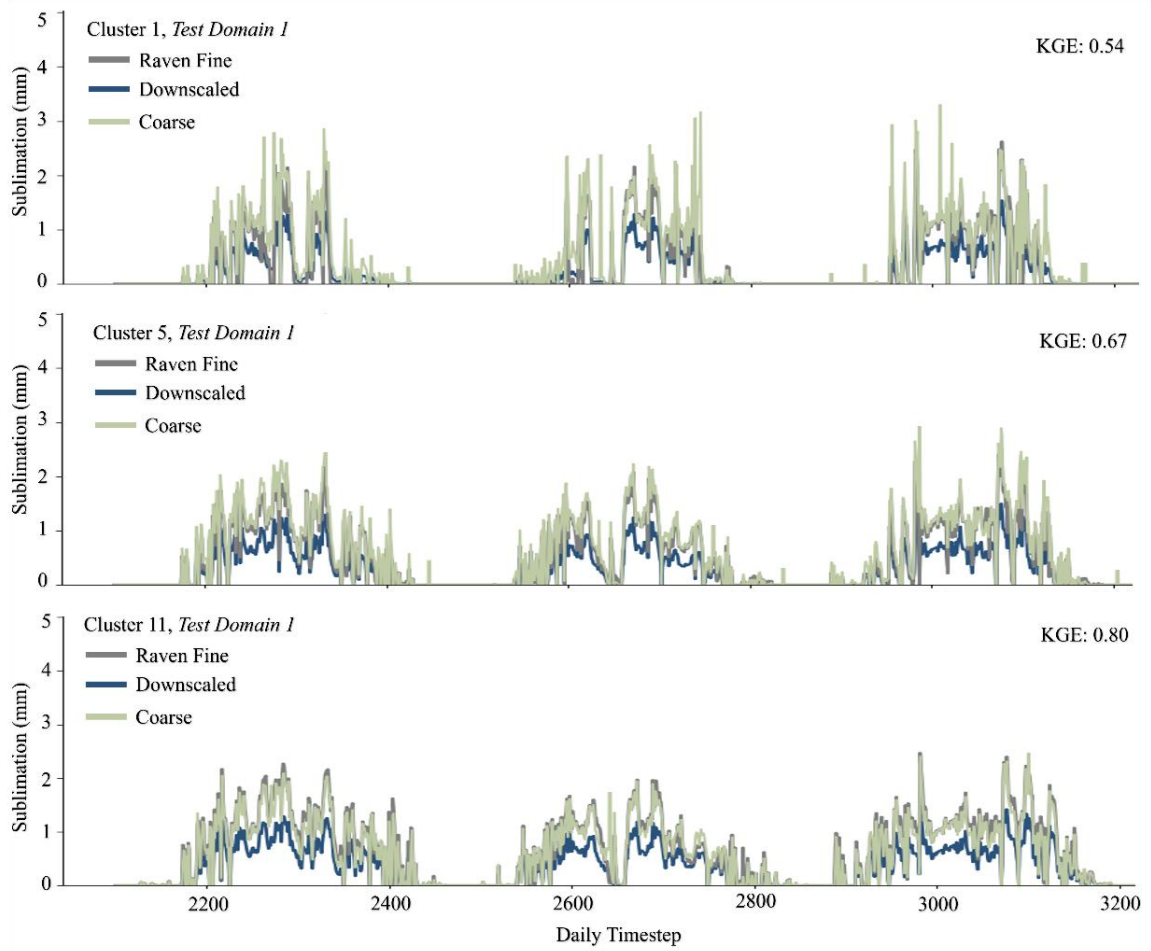


Figure 5.10: Time series plot comparing fine-resolution Raven sublimation, coarse-resolution sublimation (before downscaling), and ML-downscaled sublimation for three HRUs in clusters 1, 5, and 11.

5.3.2 Temporal Fine Scale Reconstruction

The *Temporal Validation (Ideal)* experiment (experiment 2) assessed the predictive accuracy of the ML downscaling methodology in temporal validation. The model is trained on one domain-specific dataset, utilizing 70% of the available data (14 years: 2000–2014) and estimating state scaling ratios for the remaining 30% of the domain’s dataset (6 years: 2014–2020) and this is repeated for all 5 domains. The 70/30 training split ensures a balance between data diversity for training and a sizeable evaluation

dataset. As in the *Spatial Validation (Ideal) experiment*, the domain-mean values are obtained by aggregating high-resolution estimates from the fine-scale model, rather than derived from a coarser model. This approach tests the model’s capacity to capture interannual variability in scaling ratios, with validation against the Raven fine-resolution model demonstrating its potential to extend downscaling performance across a longer temporal scope.

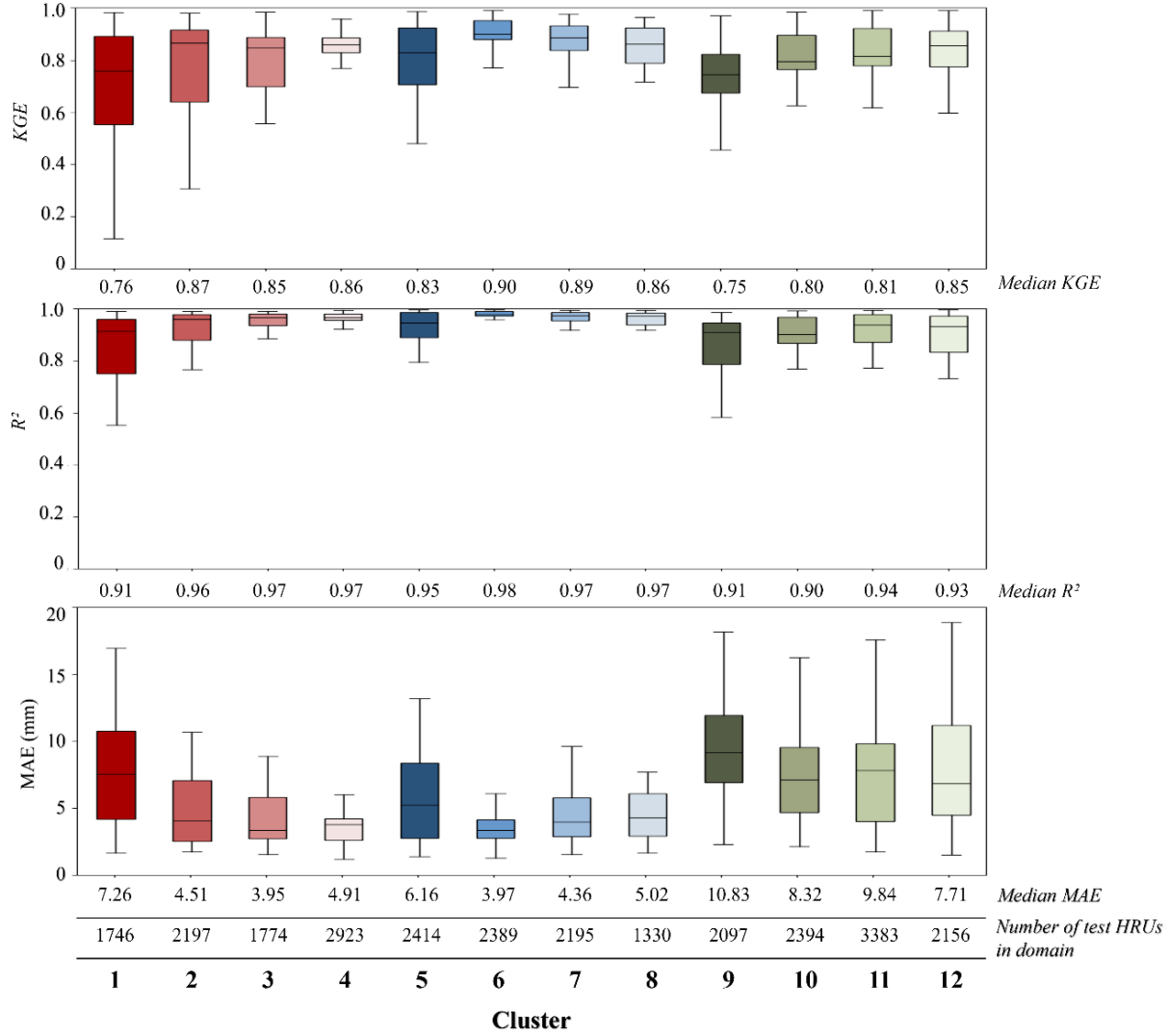


Figure 5.11: KGE, R^2 , and MAE values for the 12 clusters, evaluating the agreement between ML-downscaled SWE values and the original Raven fine-resolution estimates in the Temporal Validation (Ideal) experiment.

The results of the *Temporal Validation (Ideal) experiment* demonstrated a high level of agreement between the downscale SWE and the original fine-resolution Raven estimates. The experiment achieved KGE values comparable to the median KGE values observed in the *Spatial Validation (Ideal)* experiment, with approximately 0.8 across clusters, except for cluster 1, which exhibited notably lower

performance, as shown in Figure 5.11. The poor performance of cluster 1 across both experiments can likely be attributed to the high variability within the cluster. Although predominately forested (80.6%), cluster 1 features the most diverse landcover among the clusters including lakes (5.6%), grasslands (7.6%), shrublands (2.0%), and scattered alpine areas. Forested regions complicate snow accumulation and melt due to canopy interception, shading, and variable sub-canopy radiation, resulting in spatially inconsistent snow distribution (Essery et al., 2008; Harder and Pomeroy, 2019). Grasslands and shrublands, with less canopy cover, experience greater solar exposure, accelerating snow ablation compared to forested areas (Kremsa et al., 2015). The diverse landcover and topography within cluster 1 creates highly variable surface energy balances, increasing the complexity of snow processes and making it more challenging for the ML model to accurately reconstruct SWE.

The wider spread of R^2 values and higher observed MAE for clusters at higher elevations, indicates larger deviations from the original Raven model fine-resolution SWE estimates. Figure 5.12 shows the spatial extent of SWE values for March 7, 2019, comparing the original fine-resolution Raven SWE, the ML-downscaled SWE from the *Temporal Validation (Ideal)* experiment, and the coarse SWE that was subsequently downscaled. The light blue areas indicate regions of high SWE, while the darker purple areas represent regions with lower SWE. A KGE value of 0.89 suggest a high level of agreement between the ML-downscaled and fine-resolution SWE for March 7, 2019, indicating the model's ability to capture overall patterns. However, the zoomed-in portion of Figure 5.12 reveals that the ML-downscaled SWE values are lower in areas of high accumulation than the fine-resolution SWE estimates. This suggests that while the model was particularly successful at downscaling HRUs within low to mid elevations and capturing the general spatial heterogeneity of SWE, it struggled to fully replicate peak SWE values for HRUs at high elevations during peak snow periods.

2019-03-7

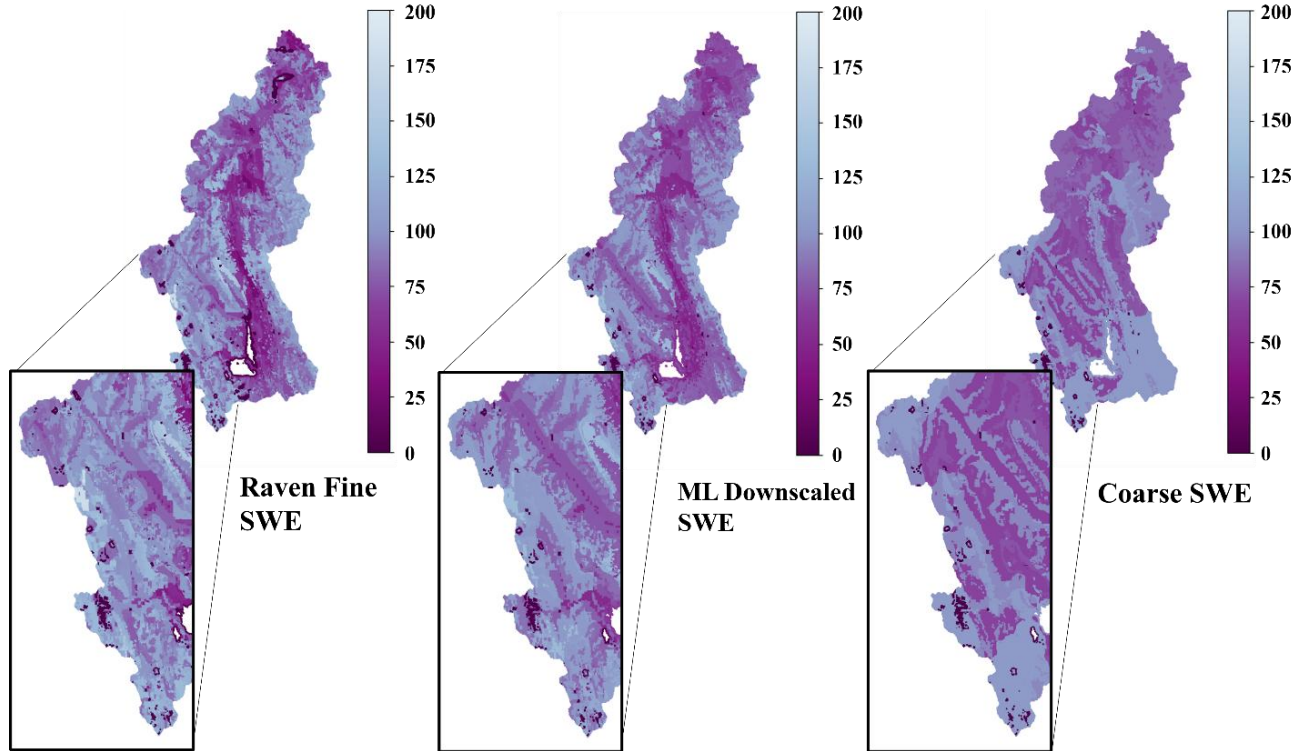


Figure 5.12: Illustrates the spatial distribution of SWE values on March 7, 2019, comparing the original fine-resolution Raven SWE, the ML-downscaled SWE from the Temporal Validation (Ideal) experiment, and the coarse SWE used for downscaling.

A similar pattern emerges during peak snowmelt periods, where the model struggled to accurately capture peak SWE values at higher elevations. Figure 5.13 shows the spatial distribution of SWE on May 30, 2019, comparing the original fine-resolution Raven SWE, the ML-downscaled SWE, and the coarse SWE used for downscaling. The ML-downscaled and fine-resolution SWE for this date achieved a KGE value of 0.72, indicating a strong overall agreement. However, the zoomed-in areas reveal that the ML-downscaled SWE tends to underestimate SWE, particularly in higher-elevation HRUs where snow persists longer. This suggests that while the downscaling approach effectively captures the overall melt pattern, it struggles to maintain accuracy in regions with extended snow retention. Despite localized underestimations, the improved representation of SWE provided by the ML-downscaled methodology offers important benefits for water management practices. By delivering finer-resolution estimates of SWE, its total volume, and its spatial distribution, the downscaled product may enable more accurate forecasting of the timing and volume snowmelt-driven runoff. Such information plays an important role in applications where local snow distribution matters such as avalanche risk assessment, managing alpine ecosystems, and supports climate change monitoring by enabling the detection of shifts in snow

accumulation and ablation patterns. Even moderate improvements in capturing spatial snow variability may strengthen water allocation strategies and emergency planning efforts.

2019-05-30

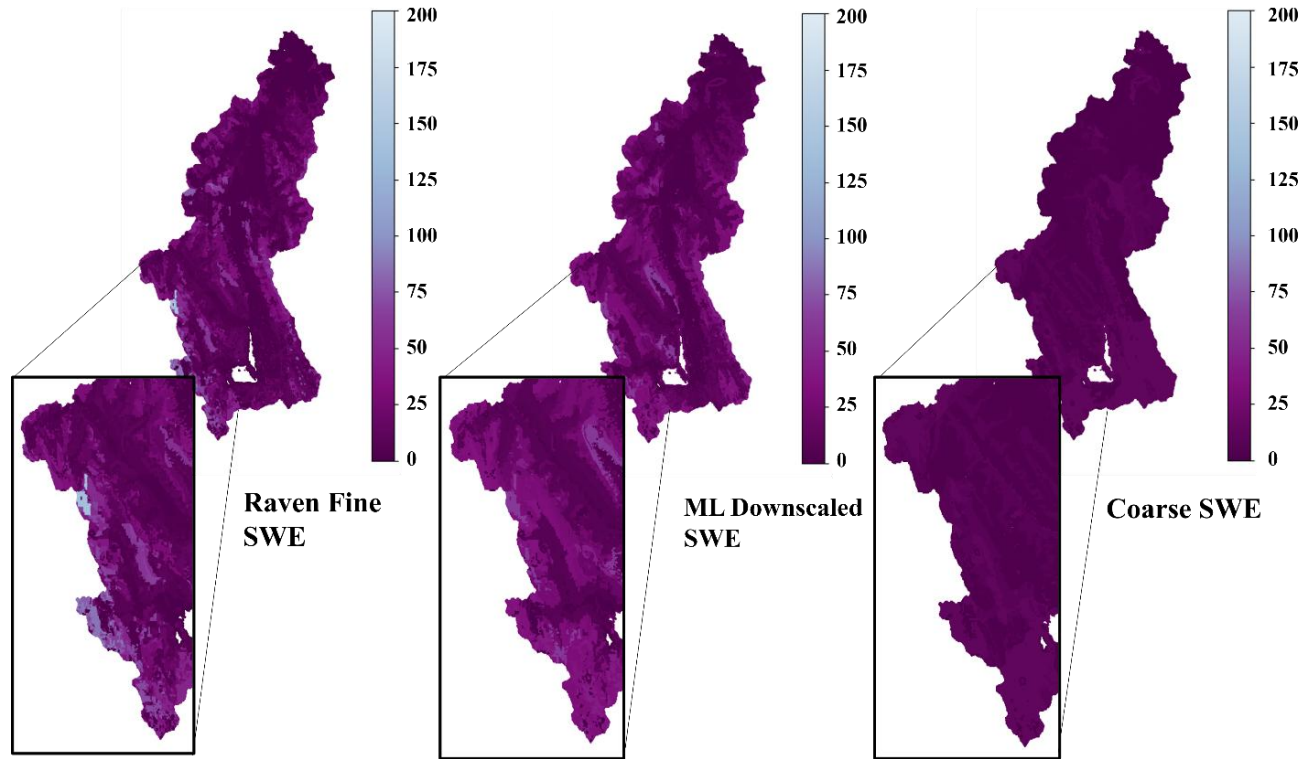


Figure 5.13: Illustrates the spatial distribution of SWE values on May 30, 2019, comparing the original fine-resolution Raven SWE, the ML-downscaled SWE from the Temporal Validation (Ideal) experiment, and the coarse SWE used for downscaling.

The strong agreement between observed and ML-derived SCA highlights the model’s ability to accurately represent snow presence and absence across clusters, a critical factor for simulating snow mass balance and energy fluxes. The SCA error metrics were comparable in magnitude to the SWE error metrics, further confirming the robustness of the ML downscaling approach. This consistency underscores the model’s ability to not only capture the spatial distribution of SWE but also accurately translate those patterns into SCA estimates. Figure 5.14 presents the SCA derived from fine-resolution Raven SWE compared to the SCA derived from the ML-downscaled SWE for May 30, 2019. Both maps exhibit a strong spatial agreement, with snow coverage predominantly confined to higher elevations, while lower elevations and valley regions are largely snow-free. The downscaled SCA effectively mirrors the fine-resolution patterns, though minor discrepancies appear in areas with patchy, residual snow cover. This capability is particularly valuable for predicting runoff timing and volume in snow-dominated watersheds, where accurate estimates of snow coverage directly influence hydrologic simulations and water resource management decisions. However, the performance of both the ML downscaling of SWE

and SCA further suggests that once established, these scaling ratios can be applied beyond the initial training dataset with moderate to high success, provided that future conditions remain reasonably similar to those of the training period.

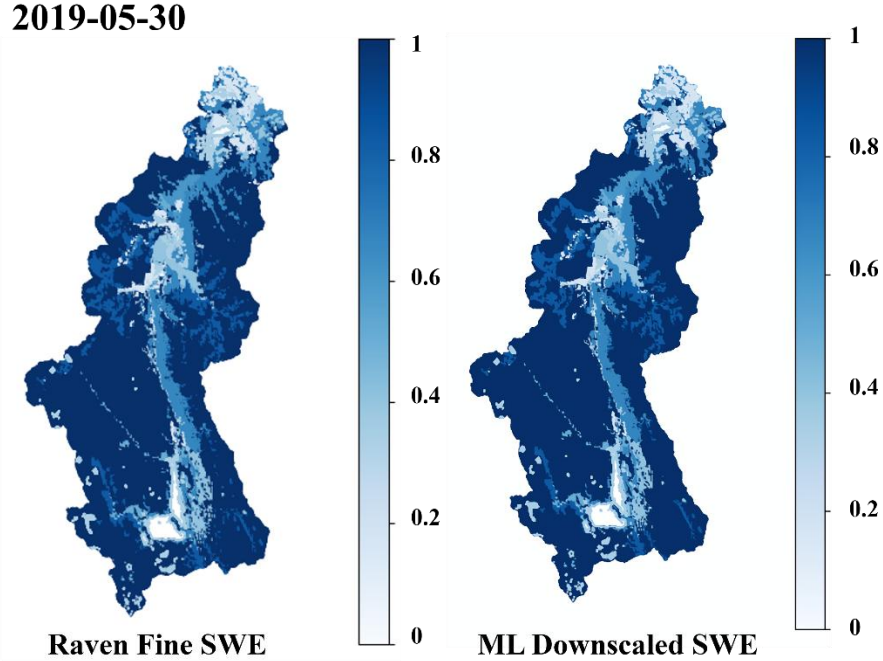


Figure 5.14: Illustrates the spatial distribution of SCA values on May 30, 2019, comparing the SCA calculated at the domain level from the original fine-resolution Raven SWE and the ML-downscaled SCA within the Temporal Validation (Ideal) experiment.

5.3.3 Validation of Upscaling and Downscaling Methodologies

The *Spatial Validation (Upscaling Input)* and *Spatial Validation (Naïve Input)* experiments both evaluate the ML downscaling methodology's ability to reconstruct fine-resolution snow state and flux model outputs using outputs from two different existing coarse-scale upscaling methods, rather than the averaged fine scale model results. First, the Chapter 3 ML upscaling method is applied to upscale SWE. This method uses mean states and forcings from training HRUs to estimate coarse-resolution fluxes and states at the (coarse) domain scale. A mass balance equation integrates these upscaled fluxes into a hydrological model to simulate coarse-scale mean states. The ML downscaling ratios derived from the *Spatial Validation (Ideal)* experiment (experiment 1) are then used to downscale these coarse-scale outputs back to the HRU resolution, where they are compared against fine-resolution Raven model outputs. As a point of comparison, a naïve snow model is also employed, which treats each cluster domain as homogenous HRUs to simulate SWE. The same ML downscaling ratios are applied to reconstruct HRU-scale outputs, which are evaluated against fine-scale Raven model results. This test is expected to

further underscore the limitations and biases of both naïve and ML upscaling methodologies while demonstrating the versatility of the downscaling when using imperfect spatial averages as input.

In the *Spatial Validation (Upscaling Input)* experiment (experiment 3), the downscaled ML upscaled SWE values, compared to the fine-resolution Raven SWE, achieved median KGE values consistently exceeding 0.75, demonstrating strong agreement across most clusters, as shown in Figure 5.15. To provide a benchmark, Figure 5.15 also includes the KGE values for coarse-resolution upscaled and naïve SWE, compared to fine-resolution Raven SWE, reinforcing the expectation that downscaling should lead to improved results. Notably, in some clusters, the downscaled SWE values from this experiment outperformed those derived from the *Spatial Validation (Ideal)* experiment, particularly clusters at higher elevations (10, 11, 12). This suggests that the ML downscaling method can work better when used to downscale upscaled model results than when used to downscale the (objectively more accurate) perfect averages of SWE at the coarse scale. This is a somewhat surprising result, implying that the downscaling technique compensates for some biases in the upscaled model. The *Spatial Validation (Naïve Input)* experiment (experiment 4) demonstrated a considerable decrease in performance compared to the other two experiments, with notable variability across clusters. This decline was particularly evident in higher-elevation clusters.

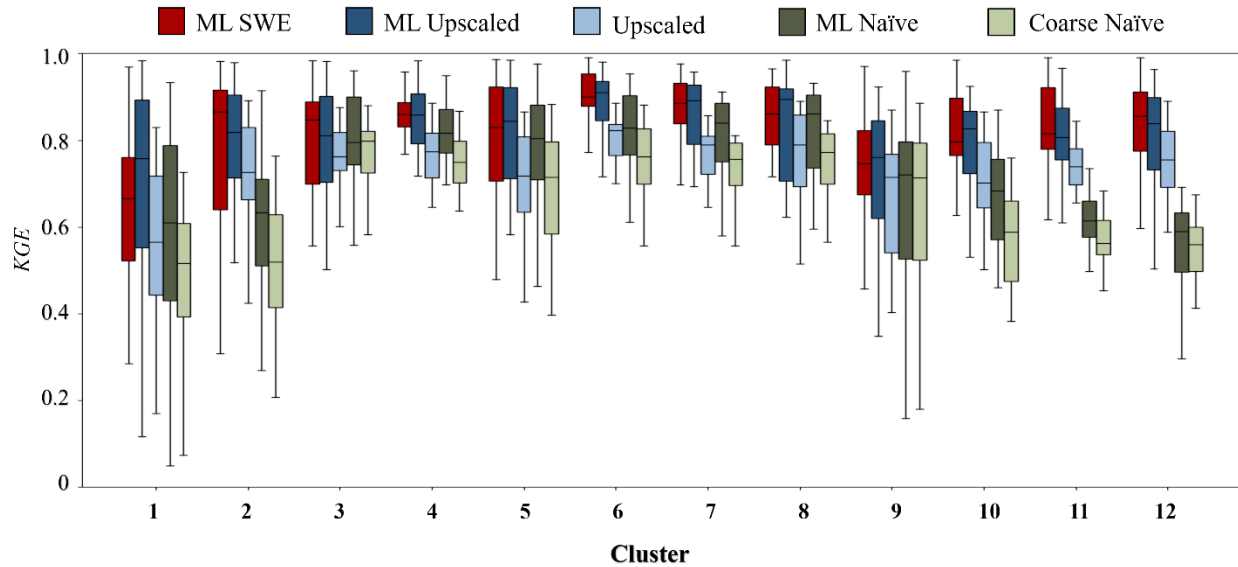


Figure 5.15: Spatial validation of KGE values for ML downscaling, ML downscaling (upscale input), coarse-resolution upsampling, ML downscaling (naïve input), and coarse-resolution naïve snow model SWE, compared against fine-resolution Raven SWE estimates.

The ML downscaling methodology showed strong potential for refining aggregated SWE estimates by preserving spatial variability typically lost during aggregation. When applied to upscaled SWE estimates, the approach effectively maintained fine-resolution topographic influences in the downscaled outputs. This capability is expected to be particularly valuable in mountainous regions, where

terrain-driven variability significantly impacts SWE accumulation and ablation patterns. It is unlikely to be as useful in milder terrain. Future work should explore the broader applicability of these ML-derived scaling ratios for downscaling coarse-resolution reanalysis or satellite-derived SWE products. Such investigations could determine whether the scaling relationships established from high-resolution training datasets hold across diverse data sources and climatic conditions.

5.4 Conclusion

This study aimed to develop an ML downscaling methodology to generate scaling ratios for reconstructing fine-resolution snow-related state and flux variables. The XGBoost-based downscaling technique demonstrated strong performance in reconstructing fine scale patterns in SWE, snowmelt, and sublimation rates from coarse resolution inputs, whether these inputs were from averages of a fine scale model (ideal) or generated by a coarse scale model. While, in spatial validation, the model effectively captured the general spatial pattern of fine-resolution SWE, it struggled to reproduce peak SWE values, particularly in the highest elevation regions. A similar pattern was observed during temporal validation. The model's difficulty in accurately downscaling SWE at higher elevations can be attributed to several factors inherent to alpine environments. These regions experience complex orographic effects, with steep gradients, variable wind redistribution, and localized accumulation patterns that are challenging to capture using coarse-scale predictors. Addressing these challenges may require incorporating additional fine-resolution atmospheric predictors to better downscale the influence of processes governing snow accumulation and ablation in complex mountain environments.

In both spatial and temporal validation of the downscaling approach, the downscaled SWE values were also used to estimate SCA and then compared to the SCA derived from the original fine-resolution model. The ML-estimated SCA effectively captured the spatial and temporal patterns of snow cover, achieving moderate to high agreement with the fine-resolution estimates. Although some discrepancies appeared during the late melt stages, when the ML-downscaled SCA struggled to capture areas of high snow retention, the strong agreement between the ML-derived and fine-resolution SCA further demonstrated the model's ability to represent the spatial heterogeneity of SWE. Spatial validation was also applied to evaluate downscaling of snowmelt and sublimation fluxes, with varying success. The model accurately represented snowmelt heterogeneity across diverse terrain conditions but struggled to reconstruct sublimation.

The ML downscaling methodology developed here is unique in that it directly downscales hydrologic model outputs, bypassing the traditional approach of first downscaling forcing variables and then integrating them into a hydrologic model to derive small-scale state and flux variables. Through the use of scaling ratios, this method eliminates the need to downscale forcing variables and run a fine-scale model to generate small-scale state and flux variables, leading to significant computational benefits.

While ML has been applied to downscaling in other contexts, this represents a novel application of ML algorithms, focusing specifically on inferring scale-dependent relationships to enhance the resolution of hydrologic model outputs. The methods ability to preserve fine-resolution details should be valuable in mountainous terrain, where steep gradients, complex topography, and localized microclimatic variations lead to high spatial variability in snow dynamics.

Chapter 6

Conclusions

6.1 Summary

Snow processes exhibit pronounced spatial heterogeneity, driven by complex energy and mass exchanges across scales. As a result, computational upscaling or downscaling of snow-related state and flux variables requires careful consideration of small-scale processes, their aggregate effects, and the relationship between fine-scale variability and coarse-scale behaviour. Although advancements in computational power suggest that finely discretized distributed models could eventually resolve many scaling challenges, scaling investigations remain valuable for understanding the factors and relationships that govern hydrological processes at larger scales. They provide critical insights into fundamental scaling relationships, making both upscaling and downscaling enduring focuses of hydrological research. They also have practical value: use of upscaled models and downscaling techniques can significantly reduce the computational costs of hydrological models; they enable simulation at the coarse scale while explicitly respecting fine-scale phenomenon. This is valuable for calibration, uncertainty analysis, and sensitivity analysis, or wherever computational cost is an issue.

This study aimed to develop, implement, and test an ML upscaling methodology for estimating spatially aggregated snow-related state and flux variables, such as SWE, snowmelt, sublimation, and SCA. Chapter 3 presents a proof of concept applied to a drainage basin in the Canadian Rockies. A fine-resolution Raven hydrologic model generated high-resolution state and flux variables, which were aggregated to train the ML model. The ML model then estimated spatially averaged results from discretized fine-scale models without detailed local response information, both with and without simulating low-order forcing statistics. Grouping HRUs with similar topographic characteristics using a k-

means clustering algorithm enhanced RF model performance, with further improvements achieved by incorporating the distributional information of forcing variables. This demonstrated the methodology's potential to preserve system heterogeneity and outperform naïve upscaling estimates when leveraging ML-generated constitutive relationships. The research outlined in Chapter 4 further evaluated the methodology's ability to identify conditions under which emergent constitutive relationships remain valid, examining ML algorithm performance, ideal spatial resolution, the minimum variable set required for comparable performance, and the transferability of relationships between study areas with similar topographic and landcover characteristics. In Chapter 5, a similar ML approach was applied to the downscaling problem, reconstructing snow-related state and flux variables back to the fine resolution using ML-estimated scaling ratios. The methodology showed moderate to high success in reconstructing fine-scale snow states using fine-resolution topographic and shortwave inputs alongside coarse-resolution forcing variables, though challenges remained in capturing peak values, with some reconstructions exhibiting additional noise. Together, the analyses in Chapters 3, 4, and 5 highlight the potential of ML to address scaling challenges in hydrology. These chapters demonstrate how ML techniques effectively capture complex relationships between topographic, landcover, and snow-related accumulation and ablation processes while enhancing computational efficiency and predictive accuracy.

6.2 Contributions

The primary contributions from this study are as follows:

- This study demonstrates the ability of ML to derive physics-informed, scale-dependent constitutive relationships between snow states, fluxes, and forcings by leveraging data from fine-resolution hydrological models.
- This is the first study to use bottom-up ML upscaling strategies to generate upscaled relations that can be used to predict instantaneous coarse-scale hydrologic fluxes in heterogeneous domains. It is expected this strategy may be extended to upscaling of other hydrologic fluxes, such as evapotranspiration or infiltration.
- By incorporating low-order statistical moments of significant forcing variables, the methodology was shown to effectively capture distributional information typically lost during aggregation. This approach outperforms simpler upscaling methods while preserving the influence of system heterogeneity.
- Variable reduction experiments identified a minimal set of key variables (e.g., SWE, shortwave radiation) that maintain predictive accuracy. This approach enhances the interpretability of ML applications, by providing insights into the variables that play significant roles in constitutive relationships.

- The transferability of the constitutive relationships derived from the Kananaskis Basin to the Oldman Basin highlight the methodology's adaptability. This capability demonstrates its potential for application to regions with similar topographic and landcover characteristics.
- The novel use of scaling ratios predicted by ML methods enables the efficient downscaling of hydrological model outputs. The downscaling strategy preserves mass balance and spatial variability, providing a robust and versatile methodology for reconstructing small-scale snow-related state and flux variables. This technique is also expected to be extendible to other hydrologic states and fluxes.

6.3 Limitations and Future Work

The proposed ML upscaling and downscaling methodologies are not without limitations. Their effectiveness relies heavily on the quality and representativeness of the fine-resolution outputs generated by the Raven hydrological model. It is essential that the model's underlying physics can capture the key processes governing snow accumulation and ablation. For example, a simple degree-day snowmelt model overlooks critical factors such as shortwave radiation or sublimation. If a simpler model such as this were used for upscaling, it would hamper the physics-informed decisions made by the ML methodology and reduce the validity (and utility) of any derived scaling relationships relative to a more physically appropriate model. This thesis intentionally did not deeply investigate the validity of the snow model used at the fine scale; it focused instead on a computational upscaling and downscaling strategy that could be applied in principle to any snow model. However, it is expected that the ability for an upscaling or downscaling approach to infer 'truth' about scaling will always be limited by the fidelity of the original model upon which it is based. Missing or simplified representation of processes at the fine scale influence how the constitutive relationships are derived, making it difficult to fully understand or verify the scaling properties captured by the models.

A primary goal of this study was to identify the key controls on upscaling, yet this was not fully achieved. While this work was able to determine which variables influenced the RF model's performance, interpreting their role in upscaling snow processes relies on the assumption that the Gini feature importance index provides meaningful insights. However, variable significance in an RF model does not necessarily indicate that the model is capturing true physical relationships, as importance measures are independent of predictive skill. As a result, certain features may appear highly significant due to data correlations or biases in the training set rather than their actual influence on physical processes. This limitation is somewhat mitigated by the Leave-One-Out (LOO) experiment, which iteratively removes variables and assesses their impact; however, it still depends on the feature importance index to determine which variables to exclude. Consequently, this assumption may impact the interpretation of key controls on snow state and flux upscaling, potentially overemphasizing less relevant features while overlooking variables that play a more significant role in shaping upscaling relationships. To improve sensitivity analysis, future approaches should operate independently of the Gini feature importance index,

instead using changes in KGE or an alternative error metric to guide variable removal and assess whether the resulting variables remain consistent with those derived using the feature importance index. Additionally, focus should not only be on identifying key variables but also on evaluating their role in upscaling. This may perhaps be done by performing LOO and similar experiments at several scales, examining the variability of relative importance with scale. Alternative feature attribution methods, such as SHAP values, may provide deeper insights into how each variable influences model predictions, while exporting and visualizing decision tree structures could further clarify how the model makes its predictions and highlight key decision pathways. Such analysis may also be revealing when applied to upscaling distinct conceptual-numerical snow models, as it is expected that different snow balance models would have different scaling behaviours.

The naïve upscaling framework used throughout this thesis represents a simplified baseline, wherein the Raven model distributes snow states and fluxes across the domain without accounting for heterogeneity in topography, land cover, or meteorological forcings. It is acknowledged that this approach represents a lower bound for performance of an upscaling strategy, especially when evaluating results at a very coarse resolution. In practice, hydrologists use 1) finer spatial discretization to avoid the artefacts associated with naïve upscaling, and 2) calibration of model parameters to compensate for these artefacts. Because of the nature of the snow models used here, there are few influential parameters to calibrate. However, alternative less-naïve benchmarks could have been used herein by discretizing by elevation band, subbasin, and land cover type (as often done in practice) rather than relying on the clustering analysis to determine the spatial extent of the coarse domains. Although incorporating this more realistic baseline may have strengthened comparisons, the naïve model provides a valuable reference point for assessing the benefits of the proposed machine learning approach. Future studies may benefit from exploration of comparisons with alternative less-naïve upscaling strategies.

In the future, the ML upscaling and downscaling methodologies presented in this thesis could be extended and applied to other hydrologic state and flux variables beyond snow-related processes, like evapotranspiration or soil moisture. For example, in Chapter 5, the downscaling methodology was briefly applied to snowmelt and sublimation beyond its initial evaluation of SWE and SCA, illustrating its potential to be transferred to other flux variables. Additionally, this thesis has been limited to basins within the Rocky Mountains, characterized by complex terrain and diverse landcover. Extending these methodologies to more homogeneous regions, such as low-relief landscapes with uniform vegetation, would provide valuable insights into their broader applicability and limitations of these scaling methods. Ultimately, these extensions would further reinforce the utility of the ML methodologies in hydrological modelling, supporting their integration into practical tools for watershed management.

While this thesis focused exclusively on daily timescales, applying the upscaling methodology to both finer (e.g., hourly) and coarser (e.g., monthly) temporal resolutions presents a promising yet challenging avenue for future research. In particular, hourly snow models introduce complexities such as

diurnal variability in shortwave radiation, which can be strongly influenced by slope and aspect. For example, east- and west-facing slopes experience solar loading at different times of day, affecting both the timing and intensity of snowmelt and sublimation, as east-facing slopes receive more intense solar radiation during the morning hours, while west-facing slopes are exposed later in the day. Although the ML-upscaling methodology is well-suited to capture non-linear interactions between topographic features and energy inputs, it would be critical that such diurnal subtleties are explicitly captured in the training data. Without this, the model may underrepresent diurnal asymmetries or fail to generalize to sub-daily applications. Conversely, upscaling at monthly timescales may smooth over important short-term dynamics, potentially missing critical feedbacks in snow accumulation and melt processes. Nonetheless, the ML methodologies flexibility suggest it could adapt to multiple temporal scales, provided the data are appropriately structured. However, users should be aware that applying the ML-based upscaling approach to different temporal resolutions may introduce new challenges.

The purpose of this thesis was to establish computational upscaling and downscaling methodologies that served as a foundation for using physics-informed and interpretable ML to bridge the gap between fine-scale understanding and large-scale hydrological applications. By leveraging interpretable ML techniques, this research provides a framework for improving the representation of snow-related state and flux variables across scales in mountainous terrain. Future studies can build on this work to address the inherent challenges of scaling other hydrological state and flux variables, advancing hydrological modelling approaches that more effectively capture the heterogeneity and complexity of real-world systems, without always having to simulate them in great detail.

A logical next step for evaluating the ML-upscaling methodology is to explore its applicability across broader geographic extents, such as extending the methodology across Canada. This could be achieved by selecting study areas within each province that reflect diverse topographic and landcover characteristics, such as coastal highlands in Newfoundland, the prairie croplands in Manitoba, or boreal forest regions in Quebec. For each of these study areas, fine-resolution hydrologic models would first be developed to simulate snow-related processes. Subsequently, a clustering algorithm could be applied to classify areas across Canada based on topographic and landcover characteristics. These clusters could then be used to guide the transfer of upscaling relationships between regions of Canada with similar, as demonstrated in the transferability experiment presented in Chapter 4. In that chapter, each cluster was subdivided into domains, coherent units composed of HRUs selected at random within the cluster, to evaluate whether relationships learned in one domain could be reliably applied to others with similar topographic and landcover characteristics. While subbasins provide a convenient spatial unit for defining domains, their boundaries are often based on hydrologic routing rather than landscape heterogeneity, which may obscure important scaling relationships. In contrast, defining domains through clustering on topographic and landcover characteristics may yield more physically meaningful groupings for upscaling applications. However, several challenges remain in applying the ML-upscaling methodology across Canada, including determining the most appropriate fine- and coarse-scale resolutions for diverse

landscapes, addressing variations in data availability, and accounting for region-specific hydrologic processes that may influence snow accumulation and ablation. These factors must be carefully considered to ensure the reliability and transferability of upscaling relationships across diverse ecozones.

Beyond its application to snow-related processes, the upscaling methodology developed in this thesis holds potential for application to broader hydrological applications. The extension of the ML upscaling methodology to other hydrologic fluxes, such as evapotranspiration, infiltration, or surface runoff, would follow the same foundational steps but would require some additional modifications. This includes, for instance, the selection of an appropriate fine-resolution model that accurately represents the relevant small-scale physics of the flux of interest. Unlike snow processes, which can be reasonably decoupled from the rest of the hydrologic cycle and modelled as an isolated process with a high degree of physical fidelity, other fluxes, particularly those driven by vegetation dynamics, are more difficult to represent at fine spatial scales and to decouple from the rest of the water cycle. Additionally, extension to other processes will be challenged by the identification of input variables for the ML model, as the performance of the upscaling relationships is limited to the information available in the training dataset. Including relevant predictors that describe both internal system states and external drivers is essential for uncovering meaningful scaling relationships. Although the RF model used through this study had a simplified architecture, more complex fluxes with less definitive drivers would likely require deeper RF architectures and additional time would be needed to tune hyperparameters appropriately. Future applications will therefore need to balance model complexity with interpretability and ensure that the fine-scale model used for training reflects the physical mechanisms of interest to preserve the utility and transferability of the learned upscaling functions.

References

- Abatzoglou, J. T., & Brown, T. J. (2012). A comparison of statistical downscaling methods suited for wildfire applications. *International journal of climatology*, 32(5), 772-780.
- Ahmed, A. N., Othman, F. B., Afan, H. A., Ibrahim, R. K., Fai, C. M., Hossain, M. S., ... & Elshafie, A. (2019). Machine learning methods for better water quality prediction. *Journal of Hydrology*, 578, 124084.
- Ahmed, S. F. (2022). Short-term electrical load demand forecasting based on LSTM and RNN deep neural networks. *Mathematical Problems in Engineering*, doi:10.1155/2022/2316474.
- Albert, R. (2011). The influence of past and future urbanization on watershed nitrogen export and hydrology dynamics in two mid-Atlantic watersheds in Fairfax, Virginia (Doctoral dissertation).
- Alberta Environment and Sustainable Resource Development. (2014). Map of public land use zones. Government of Alberta. <http://aep.alberta.ca/recreation-public-use/recreation-on-public-land/publicland-use-zones/documents/PLUZ-GreenWhite-Jul07-2014.pdf>
- Allen, R. G., Trezza, R., & Tasumi, M. (2006). Analytical integrated functions for daily solar radiation on slopes. *Agricultural and Forest Meteorology*, 139, 55-73, doi:10.1016/j.agrformet.2006.05.012.
- Ancona, M., Ceolini, E., Öztireli, C., & Gross, M. (2019). Gradient-based attribution methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, 169-191.
- Anderson, M. C., Kustas, W. P., & Norman, J. M. (2003). Upscaling and downscaling—A regional view of the soil–plant–atmosphere continuum. *Agronomy Journal*, 95(6), 1408-1423.
- Anderson, S., & Radić, V. (2022). Interpreting deep machine learning for streamflow modeling across glacial, nival, and pluvial regimes in southwestern Canada. *Frontiers in Water*, 4, 934709, doi:10.3389/frwa.2022.934709.
- Anderton, S. P., White, S. M., & Alvera, B. (2004). Evaluation of spatial variability in snow water equivalent for a high mountain catchment. *Hydrological Processes*, 18(3), 435-453.
- Andreadis, K. M., & Lettenmaier, D. P. (2006). Assimilating remotely sensed snow observations into a macroscale hydrology model. *Advances in water resources*, 29(6), 872-886.
- Archfield, S. A., Clark, M., Arheimer, B., Hay, L. E., McMillan, H., Kiang, J. E., ... & Over, T. (2015). Accelerating advances in continental domain hydrologic modeling. *Water Resources Research*, 51(12), 10078-10091.
- Arsenault, R., Martel, J. L., Brunet, F., Brissette, F., & Mai, J. (2023). Continuous streamflow prediction in ungauged basins: long short-term memory neural networks clearly outperform traditional hydrological models. *Hydrology and Earth System Sciences*, 27(1), 139-157.
- Aydin, H. E., & Iban, M. C. (2023). Predicting and analyzing flood susceptibility using boosting-based ensemble machine learning algorithms with SHapley Additive exPlanations. *Natural Hazards*, 116(3), 2957-2991.
- Barakat, N., & Bradley, A. P. (2010). Rule extraction from support vector machines: a review. *Neurocomputing*, 74(1-3), 178-190.
- Barnett, T. P., Adam, J. C., & Lettenmaier, D. P. (2005). Potential impacts of a warming climate on water availability in snow-dominated regions. *Nature*, 438(7066), 303-309.

- Barnhart, T. B., Molotch, N. P., Livneh, B., Harpold, A. A., Knowles, J. F., & Schneider, D. (2016). Snowmelt rate dictates streamflow. *Geophysical Research Letters*, 43(15), 8006-8016.
- Barnhart, T. B., Putman, A. L., Heldmyer, A. J., Rey, D. M., Hammond, J. C., Driscoll, J. M., & Sexstone, G. A. (2024). Evaluating distributed snow model resolution and meteorology parameterizations against streamflow observations: Finer is not always better. *Water Resources Research*, 60(7), e2023WR035982.
- Baudron, P., Alonso-Sarriá, F., García-Aróstegui, J. L., Cánovas-García, F., Martínez-Vicente, D., & Moreno-Brotóns, J. (2013). Identifying the origin of groundwater samples in a multi-layer aquifer system with Random Forest classification. *Journal of Hydrology*, 499, 303-315.
- Beckstead, G. R., & Veldman, W. E. (1985). Marmot Creek experimental watershed study.
- Bennett, A., & Nijssen, B. (2021). Deep learned process parameterizations provide better representations of turbulent heat fluxes in hydrologic models. *Water Resources Research*, 57, e2020WR029328, doi:10.1029/2020WR029328.
- Bennett, K. E., Miller, G., Busey, R., Chen, M., Lathrop, E. R., Dann, J. B., ... & Wulischleger, S. D. (2022). Spatial patterns of snow distribution in the sub-Arctic. *The Cryosphere*, 16(8), 3269-3293.
- Berg, J., Reynolds, D., Quéno, L., Jonas, T., Lehning, M., & Mott, R. (2024). A seasonal snowpack model forced with dynamically downscaled forcing data resolves hydrologically relevant accumulation patterns. *Frontiers in Earth Science*, 12, 1393260.
- Bergström, S., & Lindström, G. (2015). Interpretation of runoff processes in hydrological modelling—experience from the HBV approach. *Hydrological Processes*, 29(16), 3535-3545.
- Bernhardt, M., Liston, G. E., Strasser, U., Zängl, G., & Schulz, K. (2010). High resolution modelling of snow transport in complex terrain using downscaled MM5 wind fields. *The Cryosphere*, 4(1), 99-113.
- Beven, K. J. (2000). Uniqueness of place and process representations in hydrological modelling. *Hydrology and earth system sciences*, 4(2), 203-213.
- Beven, K. J., & Kirkby, M. J. (1979). A physically based, variable contributing area model of basin hydrology/Un modèle à base physique de zone d'appel variable de l'hydrologie du bassin versant. *Hydrological sciences journal*, 24(1), 43-69.
- Beven, K., & Germann, P. (2013). Macropores and water flow in soils revisited. *Water Resources Research*, 49, 3071-3092, doi:10.1029/WR018i005p01311.
- Bhasme, P., Vagadiya, J., & Bhatia, U. (2022). Enhancing predictive skills in physically-consistent way: Physics informed machine learning for hydrological processes. *Journal of Hydrology*, 615, 128618.
- Blöschl, G., & Sivapalan, M. (1995). Scale issues in hydrological modelling: A review. *Hydrological Processes*, 9, 251-290, doi:10.1002/hyp.3360090305.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.
- Broxton, P., Ehsani, M. R., & Behrangi, A. (2024). Improving mountain snowpack estimation using machine learning with Sentinel-1, the Airborne Snow Observatory, and University of Arizona snowpack data. *Earth and Space Science*, 11(3), e2023EA002964.
- Burdett, H., & Craig, J. R. (in review). Upscaling snow accumulation and ablation using machine learning. *Water Resources Research*.

- Burdett, H.. (2023). Magpie Workflow [Computer software]. GitHub. Retrieved March 4, 2025, from https://github.com/hburdett1/Magpie_Workflow
- Burnash, R. J. (1973). A generalized streamflow simulation system: Conceptual modeling for digital computers. US Department of Commerce, National Weather Service, and State of California, Department of Water Resources.
- Bürger, G., & Chen, Y. (2005). Regression-based downscaling of spatial variability for hydrologic applications. *Journal of hydrology*, 311(1-4), 299-317.
- Cartwright, K., Hopkinson, C., Kienzie, S., & Rood, S. B. (2020). Evaluation of temporal consistency of snow depth drivers of a Rocky Mountain watershed in southern Alberta. *Hydrological Processes*, 34(25), 4996-5012.
- Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE). *Geoscientific model development discussions*, 7(1), 1525-1534.
- Chakraborty, D., Başağaoğlu, H., & Winterle, J. (2021). Interpretable vs. noninterpretable machine learning models for data-driven hydro-climatological process modeling. *Expert Systems with Applications*, 170, 114498.
- Chen, L., He, Q., Liu, K., Li, J., & Jing, C. (2019). Downscaling of GRACE-derived groundwater storage based on the random forest model. *Remote Sensing*, 11, 2979, doi:10.3390/rs11242979.
- Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).
- Cheng, G., Li, X., Zhao, W., Xu, Z., Feng, Q., Xiao, S., & Xiao, H. (2014). Integrated study of the water–ecosystem–economy in the Heihe River Basin. *National Science Review*, 1, 413-428, doi:10.1093/nsr/nwu017.
- Chernos, M., MacDonald, R. J., Straker, J., Green, K., & Craig, J. R. (2022). Simulating the cumulative effects of potential open-pit mining and climate change on streamflow and water quality in a mountainous watershed. *Science of the Total Environment*, 806, 150394.
- Choi, J., Won, J., Lee, O., & Kim, S. (2021). Usefulness of global root zone soil moisture product for streamflow prediction of ungauged basins. *Remote Sensing*, 13(4), 756.
- Choudhary, K., DeCost, B., Chen, C., Jain, A., Tavazza, F., Cohn, R., ... & Wolverton, C. (2022). Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*, 8(1), 59.
- Clark, M. P., Bierkens, M. F., Samaniego, L., Woods, R. A., Uijlenhoet, R., Bennett, K. E., ... & Peters-Lidard, C. D. (2017). The evolution of process-based hydrologic models: historical challenges and the collective quest for physical realism. *Hydrology and Earth System Sciences*, 21(7), 3427-3440.
- Clark, M. P., Nijssen, B., Lundquist, J. D., Kavetski, D., Rupp, D. E., & Woods, R. A., Rasmussen, R. M. (2015). A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resources Research*, 51, 2498-2514, doi:10.1002/2015WR017200.
- Clarke, R. T. (1973). A review of some mathematical models used in hydrology, with observations on their calibration and use. *Journal of hydrology*, 19(1), 1-20.
- Craig, J. R. (2020). Raven: User's and Developer's Manual v3.0.

- Craig, J. R., Brown, G., Chlumsky, R., Jenkinson, R. W., Jost, G., Lee, K., & Tolson, B. A. (2020), Flexible watershed simulation with the Raven hydrological modelling framework. *Environmental Modelling & Software*, 129, 104728, doi:10.1016/j.envsoft.2020.104728.
- Craig, L., Bahr, J. M., & Roden, E. E. (2010). Localized zones of denitrification in a floodplain aquifer in southern Wisconsin, USA. *Hydrogeology Journal*, 18(8), 1867.
- Crumley, R. L., Bachand, C. L., & Bennett, K. E. (2024). Snow distribution patterns revisited: A physics-based and machine learning hybrid approach to snow distribution mapping in the sub-Arctic. *Water Resources Research*, 60(9), e2023WR036180.
- Dehotin, J., & Braud, I. (2008). Which spatial discretization for distributed hydrological models? Proposition of a methodology and illustration for medium to large-scale catchments. *Hydrology and Earth System Sciences*, 12(3), 769-796.
- Deng, C., Sun, P., Yin, X., Zou, J., & Wang, W. (2024). Assessment of monthly runoff simulations based on a physics-informed machine learning framework: The effect of intermediate variables in its construction. *Journal of Environmental Management*, 362, 121299.
- Dharmadasa, V., Kinnard, C., & Baraër, M. (2022). An accuracy assessment of snow depth measurements in agro-forested environments by UAV lidar. *Remote Sensing*, 14(7), 1649.
- Dierauer, J. R., Allen, D. M., & Whitfield, P. H. (2019). Snow drought risk and susceptibility in the western United States and southwestern Canada. *Water Resources Research*, 55(4), 3076-3091.
- Dingman, S. (2002), *Physical Hydrology*. Waveland Press Inc.
- Dingman, S. L. (2015). *Physical hydrology*. Waveland press.
- Doesken, N. J., Judson, A., & Center, C. C. (1997). The snow booklet: A guide to the science, climatology, and measurement of snow in the United States.
- Donald, J. R., Soulis, E. D., Kouwen, N., & Pietroniro, A. (1995). A land cover-based snow cover representation for distributed hydrologic models. *Water Resources Research*, 31(4), 995-1009.
- Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*.
- Draeger, C., Radic, V., White, R. H., & Tessema, M. A. (2023). Evaluation of ERA5 and Dynamical Downscaling for Surface Energy Balance Modeling at Mountain Glaciers in Western Canada.
- Dutra, E., Muñoz-Sabater, J., Boussetta, S., Komori, T., Hirahara, S., & Balsamo, G. (2020). Environmental lapse rate for high-resolution land surface downscaling: An application to ERA5. *Earth and Space Science*, 7(5), e2019EA000984.
- Déry, S. J., Taylor, P. A., and Xiao, J. (1998). The Thermodynamic Effects of Sublimating, Blowing Snow in the Atmospheric Boundary Layer, *Boundary-Layer Meteorol.* 89(2), 251–283.
- Esri. (2024). Subdivide Polygon (Data Management) [Software documentation]. Esri. Retrieved March 4, 2025, from <https://pro.arcgis.com>
- Essery, R., & Pomeroy, J. (2004). Vegetation and topographic control of wind-blown snow distributions in distributed and aggregated simulations for an Arctic tundra basin. *Journal of Hydrometeorology*, 5(5), 735-744.
- Essery, R., Pomeroy, J., Ellis, C., & Link, T. (2008). Modelling longwave radiation to snow beneath forest canopies using hemispherical photography or linear regression. *Hydrological Processes: An International Journal*, 22(15), 2788-2800.

- Eum, H. I., Cannon, A. J., & Murdock, T. Q. (2017). Intercomparison of multiple statistical downscaling methods: multi-criteria model selection for South Korea. *Stochastic Environmental Research and Risk Assessment*, 31, 683-703.
- Evans, J. S., Norman, A. L., & Reid, M. L. (2021). Evidence of smoke from wildland fire in surface water of an unburned watershed. *Water Resources Research*, 57(10), e2021WR030069.
- Fassnacht, S. R. (2004). Estimating Alter-shielded gauge snowfall undercatch, snowpack sublimation, and blowing snow transport at six sites in the coterminous USA. *Hydrological processes*, 18(18), 3481-3492.
- Fatichi, S., Vivoni, E. R., Ogden, F. L., Ivanov, V. Y., Mirus, B., Gochis, D., ... & Tarboton, D. (2016). An overview of current applications, challenges, and future trends in distributed process-based models in hydrology. *Journal of Hydrology*, 537, 45-60.
- Feddes, R. A., Menenti, M., Kabat, P., & Bastiaanssen, W. G. M. (1993). Is large-scale inverse modelling of unsaturated flow with areal average evaporation and surface soil moisture as estimated from remote sensing feasible? *Journal of Hydrology*, 143, 125-152, doi:10.1016/0022-1694(93)90092-N.
- Feng, D., Fang, K., & Shen, C. (2020). Enhancing streamflow forecast and extracting insights using long-short term memory networks with data integration at continental scales. *Water Resources Research*, 56(9), e2019WR026793.
- Folberth, C., Baklanov, A., Balkovič, J., Skalský, R., Khabarov, N., & Obersteiner, M. (2019). Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning. *Agricultural and forest meteorology*, 264, 1-15.
- Fowler, H. J., Blenkinsop, S., & Tebaldi, C. (2007). Linking climate change modelling to impacts studies.
- Freeze, R. A., & Harlan, R. L. (1969). Blueprint for a physically-based, digitally-simulated hydrologic response model. *Journal of hydrology*, 9(3), 237-258.
- Gao, H., Sabo, J. L., Chen, X., Liu, Z., Yang, Z., Ren, Z., & Liu, M. (2018). Landscape heterogeneity and hydrological processes: a review of landscape-based hydrological models. *Landscape ecology*, 33, 1461-1480.
- Ge, Y., Jin, Y., Stein, A., Chen, Y., Wang, J., Wang, J., ... & Atkinson, P. M. (2019). Principles and methods of scaling geospatial Earth science data. *Earth-Science Reviews*, 197, 102897.
- Geng, J., Fan, J., Wang, H., Ma, X., Li, B., & Chen, F. (2015). High-resolution SAR image classification via deep convolutional autoencoders. *IEEE Geoscience and Remote Sensing Letters*, 12(11), 2351-2355.
- Ghobadi, F., & Kang, D. (2023). Application of machine learning in water resources management: A systematic literature review. *Water*, 15, 620, doi:10.3390/w15040620.
- Gholami, V., & Sahour, H. (2022). Simulation of rainfall-runoff process using an artificial neural network (ANN) and field plots data. *Theoretical and Applied Climatology*, 147(1), 87-98.
- Ghosh, S., & Mujumdar, P. P. (2008). Statistical downscaling of GCM simulations to streamflow using relevance vector machine. *Advances in water resources*, 31(1), 132-146.
- Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. In 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA), 80-89, doi:10.1109/DSAA.2018.00018.

- Godio, A., & Rege, R. B. (2016). Analysis of georadar data to estimate the snow depth distribution. *Journal of Applied Geophysics*, 129, 92-100.
- Golden, H. E., Creed, I. F., Ali, G., Basu, N. B., Neff, B. P., Rains, M. C., ... & Lang, M. (2017). Integrating geographically isolated wetlands into land management decisions. *Frontiers in Ecology and the Environment*, 15(6), 319-327.
- Golding, D. L., & Swanson, R. H. (1986). Snow distribution patterns in clearings and adjacent forest. *Water Resources Research*, 22(13), 1931-1940.
- Govender, P., & Sivakumar, V. (2020). Application of k-means and hierarchical clustering techniques for analysis of air pollution: A review (1980–2019). *Atmospheric Pollution Research*, 11, 40-56, doi:10.1016/j.apr.2019.09.009.
- Graf, R., Kolerski, T., & Zhu, S. (2022). Predicting ice phenomena in a river using the artificial neural network and extreme gradient boosting. *Resources*, 11(2), 12.
- Gray, D. M., & Male, D. H. (Eds.). (1981). *Handbook of snow* (Vol. 776). Toronto, Canada: Pergamon Press.
- Grayson, R., & Blöschl, G. (2000). Spatial processes, organisation and patterns. In *Spatial patterns in catchment hydrology: observations and modelling* (pp. 3-16). Cambridge: Cambridge University Press.
- Gregorutti, B., Michel, B., & Saint-Pierre, P. (2017). Correlation and variable importance in random forests. *Statistics and Computing*, 27, 659-678.
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5), 1-42.
- Gupta, A., Reverdy, A., Cohard, J. M., Hector, B., Descloitres, M., Vandervaere, J. P., ... & Voisin, D. (2023). Impact of distributed meteorological forcing on simulated snow cover and hydrological fluxes over a mid-elevation alpine micro-scale catchment. *Hydrology and Earth System Sciences*, 27(1), 191-212.
- 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, doi:10.1016/j.jhydrol.2009.08.003.
- Gupta, H. V., Sorooshian, S., & Yapo, P. O. (1999). Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration. *Journal of Hydrologic Engineering*, 4, 135-143, doi:10.1061/(ASCE)1084-0699(1999)4:2(135).
- Han, H., Yang, M., Liu, X., Li, Y., Gao, G., & Wang, E. (2023). Study on the constitutive equation and mechanical properties of natural snow under step loading. *Water*, 15(18), 3271.
- Han, M., Shen, H., Tolson, B. A., Craig, J. R., Mai, J., Lin, S. G., & Awol, F. S. (2023). BasinMaker 3.0: A GIS toolbox for distributed watershed delineation of complex lake-river routing networks. *Environmental Modelling & Software*, 164, 105688, doi:10.1016/j.envsoft.2023.105688.
- Han, M., Shen, H., Tolson, B. A., Craig, J. R., Mai, J., Lin, S. G., ... & Awol, F. S. (2023). BasinMaker 3.0: A GIS toolbox for distributed watershed delineation of complex lake-river routing networks. *Environmental Modelling & Software*, 164, 105688.
- Han, M., Shen, H., Tolson, B.A., Craig, J.R., Mai, J., Lin, S., Basu, N., Awol, F., (2021). North American Lake-River Routing Product V 2.1. <https://doi.org/10.5281/ZENODO.4728185> derived by BasinMaker GIS Toolbox

- Han, Y., Wu, J., Zhai, B., Pan, Y., Huang, G., Wu, L., & Zeng, W. (2019). Coupling a bat algorithm with xgboost to estimate reference evapotranspiration in the arid and semiarid regions of china. *Advances in Meteorology*, 2019(1), 9575782.
- Harder, P., Pomeroy, J., & Helgason, W. D. (2019). Advances in mapping sub-canopy snow depth with unmanned aerial vehicles using structure from motion and lidar techniques.
- Harter, T., Hopmans, J. W., & Feddes, R. A. (2004). Role of vadose zone flow processes in regional scale hydrology: Review, opportunities and challenges (pp. 179-208). Kluwer.
- Helbig, N., Mott, R., Bühler, Y., Le Toumelin, L., & Lehning, M. (2024). Snowfall deposition in mountainous terrain: a statistical downscaling scheme from high-resolution model data on simulated topographies. *Frontiers in Earth Science*, 11, 1308269.
- Herath, M., Jayatilaka, T., Hoshino, Y., & Rathnayake, U. (2023). Deep machine learning-based water level prediction model for Colombo flood detention area. *Applied Sciences*, 13, 2194, doi:10.3390/app13042194.
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... & Thépaut, J. N. (2020). The ERA5 global reanalysis. *Quarterly journal of the royal meteorological society*, 146(730), 1999-2049.
- Hidalgo, H. G., Dettinger, M. D., & Cayan, D. R. (2008). Downscaling with constructed analogues: Daily precipitation and temperature fields over the United States. California Energy Commission PIER Final Project Report CEC-500-2007-123.
- Hipsey, M. R., Hamilton, D. P., Hanson, P. C., Carey, C. C., Coletti, J. Z., Read, J. S., ... & Brookes, J. D. (2015). Predicting the resilience and recovery of aquatic systems: A framework for model evolution within environmental observatories. *Water Resources Research*, 51(9), 7023-7043.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8), 1735-1780.
- Hollister, J., Shah, T., Robitaille, A. L., Beck, M. W., & Johnson, M. (2017). elevatr: access elevation data from various APIs. R package version 0.1, 3.
- Hunt, K. M., Matthews, G. R., Pappenberger, F., & Prudhomme, C. (2022). Using a long short-term memory (LSTM) neural network to boost river streamflow forecasts over the western United States. *Hydrology and Earth System Sciences*, 26(21), 5449-5472.
- Hutengs, C., & Vohland, M. (2016). Downscaling land surface temperatures at regional scales with random forest regression. *Remote Sensing of Environment*, 178, 127-141.
- Huth, R. (2002). Statistical downscaling of daily temperature in central Europe. *Journal of Climate*, 15(13), 1731-1742.
- Hwang, S., Graham, W. D., Adams, A., & Geurink, J. (2013). Assessment of the utility of dynamically-downscaled regional reanalysis data to predict streamflow in west central Florida using an integrated hydrologic model. *Regional Environmental Change*, 13, 69-80.
- Iseri, Y., Diaz, A. J., Trinh, T., Kavvas, M. L., Ishida, K., Anderson, M. L., ... & Snider, E. D. (2021). Dynamical downscaling of global reanalysis data for high-resolution spatial modeling of snow accumulation/melting at the central/southern Sierra Nevada watersheds. *Journal of Hydrology*, 598, 126445.
- Jana, R. B., Mohanty, B. P., & Sheng, Z. (2012). Upscaling soil hydraulic parameters in the Picacho Mountain region using Bayesian neural networks. *Transactions of the ASABE*, 55(2), 463-473.

- Ji, H., Chen, Y., Fang, G., Li, Z., Duan, W., & Zhang, Q. (2021), Adaptability of machine learning methods and hydrological models to discharge simulations in data-sparse glaciated watersheds. *Journal of Arid Land*, 13, 549-567, doi:10.1007/s40333-021-0066-5.
- Jiang, S., Zheng, Y., & Solomatine, D. (2020). Improving AI system awareness of geoscience knowledge: Symbiotic integration of physical approaches and deep learning. *Geophysical Research Letters*, 47(13), e2020GL088229.
- Kampf, S. K., & Burges, S. J. (2007), A framework for classifying and comparing distributed hillslope and catchment hydrologic models. *Water Resources Research*, 43, doi:10.1029/2006WR005370.
- Kashinath, K., Mustafa, M., Albert, A., Wu, J. L., Jiang, C., Esmaeilzadeh, S., ... & Prabhat, N. (2021). Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A*, 379(2194), 20200093.
- Kavvas, M. L. (2003), Nonlinear hydrologic processes: Conservation equations for determining their means and probability distributions. *Journal of Hydrologic Engineering*, 8, 44-53, doi:10.1061/(ASCE)1084-0699(2003)8:2(44).
- Kavvas, M. L., Kure, S., Chen, Z. Q., Ohara, N., & Jang, S. (2013). WEHY-HCM for modeling interactive atmospheric hydrologic processes at watershed scale. I: Model description. *Journal of Hydrologic Engineering*, 18(10), 1262-1271.
- Ke, Y., Im, J., Park, S., & Gong, H. (2016). Downscaling of MODIS one kilometer evapotranspiration using Landsat-8 data and machine learning approaches. *Remote Sensing*, 8(3), 215.
- King, M. R. (2020). One approach for downscaling climate change data towards regional implications in climate change scenarios: the case for Newfoundland and Labrador, Canada. *SN Applied Sciences*, 2(12), 2039.
- Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water resources research*, 42(3).
- Kleinberg, E. (2000), On the algorithmic implementation of stochastic discrimination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22, 473–490, doi:10.1109/34.857004.
- Klemeš, V. (1983), Conceptualization and scale in hydrology. *Journal of Hydrology*, 65, 1-23, doi:10.1016/0022-1694(83)90208-1.
- Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using long short-term memory (LSTM) networks. *Hydrology and Earth System Sciences*, 22(11), 6005-6022.
- Kratzert, F., Klotz, D., Hochreiter, S., & Nearing, G. S. (2021). A note on leveraging synergy in multiple meteorological data sets with deep learning for rainfall–runoff modeling. *Hydrology and Earth System Sciences*, 25(5), 2685-2703.
- Kremsa, J., Křeček, J., & Kubin, E. (2015). Comparing the impacts of mature spruce forests and grasslands on snow melt, water resource recharge, and run-off in the northern boreal environment. *International Soil and Water Conservation Research*, 3(1), 50-56.
- Kumar, V., Sharma, K. V., Caloiero, T., Mehta, D. J., & Singh, K. (2023). Comprehensive overview of flood modeling approaches: A review of recent advances. *Hydrology*, 10(7), 141.
- Kusch, E., & Davy, R. (2022), KrigR—a tool for downloading and statistically downscaling climate reanalysis data. *Environmental Research Letters*, 17, 024005, doi:10.1088/1748-9326/ac48b3.

- Kutchment, L., & Gelfan, A. N. (1996), The determination of the snowmelt rate and the meltwater outflow from a snowpack for modelling river runoff generation. *Journal of Hydrology*, 179, 23-26, doi:10.1016/0022-1694(95)02878-1.
- Kuzmin, P. P. (1972), Melting of snow cover. Technical Report, Israel Program for Scientific Translations, Jerusalem.
- Lange, H., & Sippel, S. (2020). Machine learning applications in hydrology. *Forest-water interactions*, 233-257.
- Leiva, R. G., Anta, A. F., Mancuso, V., & Casari, P. (2019). A novel hyperparameter-free approach to decision tree construction that avoids overfitting by design. *Ieee Access*, 7, 99978-99987.
- Li, H. Y., He, Y. Q., Hao, X. H., Che, T., Wang, J., & Huang, X. D. (2015). Downscaling snow cover fraction data in mountainous regions based on simulated inhomogeneous snow ablation. *Remote Sensing*, 7(7), 8995-9019.
- Li, H., Zhang, C., Chu, W., Shen, D., & Li, R. (2024). A process-driven deep learning hydrological model for daily rainfall-runoff simulation. *Journal of Hydrology*, 637, 131434.
- Li, L., & Pomeroy, J. W. (1997). Probability of occurrence of blowing snow. *Journal of Geophysical Research: Atmospheres*, 102(D18), 21955-21964.
- Li, R., Wang, S. Y., & Gillies, R. R. (2016). A combined dynamical and statistical downscaling technique to reduce biases in climate projections: an example for winter precipitation and snowpack in the western United States. *Theoretical and Applied Climatology*, 124, 281-289.
- Likas, A., Vlassis, N., & Verbeek, J. J. (2003), The global k-means clustering algorithm. *Pattern Recognition*, 36, 451-461, doi:10.1016/S0031-3203(02)00060-2.
- Lin, H. A., Lin, L. T., & Lin, S. F. (2023). Application of artificial neural network models to differentiate between complicated and uncomplicated acute appendicitis. *Journal of Medical Systems*, 47(1), 38.
- Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. (2017). An ensemble random forest algorithm for insurance big data analysis. *Ieee access*, 5, 16568-16575.
- Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S. (1997). Development and test of the distributed HBV-96 hydrological model. *Journal of hydrology*, 201(1-4), 272-288.
- Liston, G. E. (1999). Interrelationships among snow distribution, snowmelt, and snow cover depletion: Implications for atmospheric, hydrologic, and ecologic modeling. *Journal of Applied Meteorology and Climatology*, 38(10), 1474-1487.
- Liston, G. E. (2004). Representing subgrid snow cover heterogeneities in regional and global models. *Journal of climate*, 17(6), 1381-1397.
- Liston, G. E., & Sturm, M. (1998). A snow-transport model for complex terrain. *Journal of Glaciology*, 44(148), 498-516.
- Litaor, M. I., Williams, M., & Seastedt, T. R. (2008). Topographic controls on snow distribution, soil moisture, and species diversity of herbaceous alpine vegetation, Niwot Ridge, Colorado. *Journal of Geophysical Research: Biogeosciences*, 113(G2).
- Liu, J., Cao, F., Gao, X. Z., Yu, L., & Liang, J. (2020, April). A cluster-weighted kernel k-means method for multi-view clustering. In *Proceedings of the aaai conference on artificial intelligence* (Vol. 34, No. 04, pp. 4860-4867).

- Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. *Advances in neural information processing systems*, 26.
- 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(6), 1421-1438.
- Luce, C. H., Tarboton, D. G., & Cooley, K. R. (1999). Sub-grid parameterization of snow distribution for an energy and mass balance snow cover model. *Hydrological Processes*, 13(12-13), 1921-1933.
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Lundquist, J. D., & Dettinger, M. D. (2005). How snowpack heterogeneity affects diurnal streamflow timing. *Water Resources Research*, 41(5).
- Lundquist, J., Abel, M. R., Gutmann, E., & Kapnick, S. (2019). Our skill in modeling mountain rain and snow is bypassing the skill of our observational networks. *Bulletin of the American Meteorological Society*, 100(12), 2473-2490.
- Lv, Z., Pomeroy, J. W., & Fang, X. (2019). Evaluation of SNODAS snow water equivalent in western Canada and assimilation into a Cold Region Hydrological Model. *Water Resources Research*, 55(12), 11166-11187.
- López-Moreno, J. I., Revuelto, J., Gilaberte, M., Morán-Tejeda, E., Pons, M., Jover, E., ... & Pomeroy, J. W. (2014). The effect of slope aspect on the response of snowpack to climate warming in the Pyrenees. *Theoretical and Applied Climatology*, 117, 207-219.
- MacQueen, J. (1967, January). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics (Vol. 5, pp. 281-298)*. University of California press.
- Mao, G., Wang, M., Liu, J., Wang, Z., Wang, K., Meng, Y., ... & Li, Y. (2021). Comprehensive comparison of artificial neural networks and long short-term memory networks for rainfall-runoff simulation. *Physics and Chemistry of the Earth, Parts a/b/c*, 123, 103026.
- Marsh, C. B., Lv, Z., Vionnet, V., Harder, P., Spiteri, R. J., & Pomeroy, J. W. (2024). Snowdrift-permitting simulations of seasonal snowpack processes over large mountain extents. *Water Resources Research*, 60(8), e2023WR036948.
- Marsh, C. B., Vionnet, V., & Pomeroy, J. W. (2023). Windmapper: An efficient wind downscaling method for hydrological models. *Water Resources Research*, 59(3), e2022WR032683.
- McDonnell, J. J., Sivapalan, M., Vaché, K., Dunn, S., Grant, G., Haggerty, R., ... & Weiler, M. (2007). Moving beyond heterogeneity and process complexity: A new vision for watershed hydrology. *Water Resources Research*, 43(7).
- McKay, G. A., & Gray, D. M. (1981). The distribution of snowcover. *Handbook of snow*, 153, 190.
- Meloche, J., Langlois, A., Rutter, N., McLennan, D., Royer, A., Billecocoq, P., & Ponomarenko, S. (2022). High-resolution snow depth prediction using random forest algorithm with topographic parameters: A case study in the Greiner watershed, Nunavut. *Hydrological Processes*, 36(3), e14546.
- Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., & Hamprecht, F. A. (2009). A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC bioinformatics*, 10, 1-16.

- Mishra, A., & Mahajan, P. (2004). A constitutive law for snow taking into account the compressibility. *Annals of Glaciology*, 38, 145-149.
- Mital, U., Dwivedi, D., Brown, J. B., & Steefel, C. I. (2022). Downscaled hyper-resolution (400 m) gridded datasets of daily precipitation and temperature (2008–2019) for the East–Taylor subbasin (western United States). *Earth System Science Data*, 14(11), 4949-4966.
- Mital, U., Dwivedi, D., Özgen-Xian, I., Brown, J. B., & Steefel, C. I. (2022). Modeling spatial distribution of snow water equivalent by combining meteorological and satellite data with lidar maps. *Artificial Intelligence for the Earth Systems*, 1(4), e220010.
- Moghaddam, M. A., Ferre, P. A., Ehsani, M. R., Klakovich, J., & Gupta, H. V. (2021). Can deep learning extract useful information about energy dissipation and effective hydraulic conductivity from gridded conductivity fields? *Water*, 13, 1668, doi:10.3390/w13121668.
- Moore, R. J. (1985). The probability-distributed principle and runoff production at point and basin scales. *Hydrological Sciences Journal*, 30(2), 273-297.
- Moradzadeh, M., Alijanian, M., & Moeini, R. (2023). Spatial Downscaling of Snow Water Equivalent Using Machine Learning Methods Over the Zayandehroud River Basin, Iran. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 91(5), 391-404.
- Moreno, J. I. L., Latron, J., & Lehmann, A. (2010). Effects of sample and grid size on the accuracy and stability of regression-based snow interpolation methods. *Hydrological Processes*, 24(14), 1914-1928.
- Morin, S., Horton, S., Techel, F., Bavay, M., Coléou, C., Fierz, C., ... & Vionnet, V. (2020). Application of physical snowpack models in support of operational avalanche hazard forecasting: A status report on current implementations and prospects for the future. *Cold regions science and technology*, 170, 102910.
- Mushtaq, H., Akhtar, T., Hashmi, M. Z. U. R., Masood, A., & Saeed, F. (2024). Hydrologic interpretation of machine learning models for 10-daily streamflow simulation in climate sensitive upper Indus catchments. *Theoretical and Applied Climatology*, 155(6), 5525-5542.
- Najafi, M. R., Moradkhani, H., & Wherry, S. (2010, November). Optimal selection of GCM predictors combined with a new procedure for statistical downscaling of precipitation. In *World Environmental and Water Resources Congress 2010: Challenges of Change* (pp. 4598-4607).
- Natural Resources Canada (NRCAN) 2016. Canadian Digital Elevation Model. 2016-10-21. Government of Canada, Sherbrooke, Quebec.
- Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., ... & Gupta, H. V. (2021). What role does hydrological science play in the age of machine learning?. *Water Resources Research*, 57(3), e2020WR028091.
- Neuman, S. P., & Di Federico, V. (1998). Correlation, flow, and transport in multiscale permeability fields. *Scale dependence and scale invariance in hydrology*, 354.
- Ng, K. W., Huang, Y. F., Koo, C. H., Chong, K. L., El-Shafie, A., & Ahmed, A. N. (2023). A review of hybrid deep learning applications for streamflow forecasting. *Journal of Hydrology*, 625, 130141.
- Nguyen, V. T. V., Nguyen, T. D., & Gachon, P. (2006). On the linkage of large-scale climate variability with local characteristics of daily precipitation and temperature extremes: an evaluation of statistical downscaling methods. In *Advances in Geosciences: Volume 4: Hydrological Science (HS)* (pp. 1-9).

- Ohara, N., Kavvas, M. L., & Chen, Z. Q. (2008), Stochastic upscaling for snow accumulation and melt processes with PDF approach. *Journal of Hydrologic Engineering*, 13, 1103-1118, doi:10.1061/(ASCE)1084-0699(2008)13:12(1103).
- Ohara, N., Kavvas, M. L., Chen, Z. Q., Liang, L., Anderson, M., Wilcox, J., & Mink, L. (2014), Modelling atmospheric and hydrologic processes for assessment of meadow restoration impact on flow and sediment in a sparsely gauged California watershed. *Hydrological Processes*, 28, 3053-3066, doi:10.1002/hyp.9821.
- Olsson, J., Berggren, K., Olofsson, M., & Viklander, M. (2009). Applying climate model precipitation scenarios for urban hydrological assessment: A case study in Kalmar City, Sweden. *Atmospheric Research*, 92(3), 364-375.
- Padrón, R. S., Gudmundsson, L., Greve, P., & Seneviratne, S. I. (2017). Large-scale controls of the surface water balance over land: Insights from a systematic review and meta-analysis. *Water Resources Research*, 53(11), 9659-9678.
- Pang, B., Yue, J., Zhao, G., & Xu, Z. (2017), Statistical downscaling of temperature with the random forest model. *Advances in Meteorology*, 2017, doi:10.1155/2017/7265178.
- Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of hydrology*, 279(1-4), 275-289.
- Pflug, J. M., Hughes, M., & Lundquist, J. D. (2021). Downscaling snow deposition using historic snow depth patterns: Diagnosing limitations from snowfall biases, winter snow losses, and interannual snow pattern repeatability. *Water Resources Research*, 57(8), e2021WR029999.
- Pomeroy, J. W., & Gray, D. M. (1995). Snowcover accumulation, relocation and management (NHRI Science report No. 7) National Hydrology Research Institute. Environment Canada, Saskatoon (144pp).
- Pomeroy, J. W., Gray, D. M., Shook, K. R., Toth, B., Essery, R. L. H., Pietroniro, A., & Hedstrom, N. (1998). An evaluation of snow accumulation and ablation processes for land surface modelling. *Hydrological Processes*, 12(15), 2339-2367.
- Pomeroy, J. W., Parviainen, J., Hedstrom, N., & Gray, D. M. (1998). Coupled modelling of forest snow interception and sublimation. *Hydrological processes*, 12(15), 2317-2337.
- Pomeroy, J., Essery, R., & Toth, B. (2004). Implications of spatial distributions of snow mass and melt rate for snow-cover depletion: observations in a subarctic mountain catchment. *Annals of Glaciology*, 38, 195-201.
- Praskievicz, S. (2018). Downscaling climate-model output in mountainous terrain using local topographic lapse rates for hydrologic modeling of climate-change impacts. *Physical Geography*, 39(2), 99-117.
- Qi, W. Y., Chen, J., Li, L., Xu, C. Y., Li, J., Xiang, Y., & Zhang, S. (2022). Regionalization of catchment hydrological model parameters for global water resources simulations. *Hydrology Research*, 53(3), 441-466.
- Raghavendra, N., & Deka, P. C. (2015). Forecasting monthly groundwater level fluctuations in coastal aquifers using hybrid Wavelet packet–Support vector regression. *Cogent Engineering*, 2(1), 999414.
- Rahman, A. S., Hosono, T., Quilty, J. M., Das, J., & Basak, A. (2020), Multiscale groundwater level forecasting: Coupling new machine learning approaches with wavelet transforms. *Advances in Water Resources*, 141, 103595, doi:10.1016/j.advwatres.2020.103595.

- Raleigh, M. S., Lundquist, J. D., & Clark, M. P. (2015). Exploring the impact of forcing error characteristics on physically based snow simulations within a global sensitivity analysis framework. *Hydrology and Earth System Sciences*, 19(7), 3153-3179.
- Ranjram, M. (2021), Upscaling of lateral groundwater flow processes in watershed models.
- Refsgaard, J. C. (1997). Validation and intercomparison of different updating procedures for real-time forecasting. *Hydrology Research*, 28(2), 65-84.
- Ren, H., Pang, B., Bai, P., Zhao, G., Liu, S., Liu, Y., & Li, M. (2024). Flood susceptibility assessment with random sampling strategy in ensemble learning (RF and XGBoost). *Remote Sensing*, 16(2), 320.
- Ren, W., Li, X., Zheng, D., Zeng, R., Su, J., Mu, T., & Wang, Y. (2023). Enhancing flood simulation in data-limited glacial river basins through hybrid modeling and multi-source remote sensing data. *Remote Sensing*, 15(18), 4527.
- Revuelto, J., Alonso-González, E., Gascoin, S., Rodríguez-López, G., & López-Moreno, J. I. (2021). Spatial downscaling of MODIS snow cover observations using Sentinel-2 snow products. *Remote Sensing*, 13(22), 4513.
- Revuelto, J., Billecocoq, P., Tuzet, F., Cluzet, B., Lamare, M., Larue, F., & Dumont, M. (2020). Random forests as a tool to understand the snow depth distribution and its evolution in mountain areas. *Hydrological Processes*, 34(26), 5384-5401.
- Rokach, L., & Maimon, O. (2005). Decision trees. *Data mining and knowledge discovery handbook*, 165-192.
- Réveillet, M., MacDonell, S., Gascoin, S., Kinnard, C., Lhermitte, S., & Schaffer, N. (2020). Impact of forcing on sublimation simulations for a high mountain catchment in the semiarid Andes. *The Cryosphere*, 14(1), 147-163.
- 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, doi:10.1016/j.jhydrol.2023.130380.
- Sahoo, B. B., Jha, R., Singh, A., & Kumar, D. (2019). Long short-term memory (LSTM) recurrent neural network for low-flow hydrological time series forecasting. *Acta Geophysica*, 67(5), 1471-1481.
- Samaniego, L., Kumar, R., & Attinger, S. (2010), Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resources Research*, 46, doi:10.1029/2008WR007327.
- Saydi, M., & Ding, J. L. (2020). Impacts of topographic factors on regional snow cover characteristics. *Water Science and Engineering*, 13(3), 171-180.
- Schmidt, L., Heße, F., Attinger, S., & Kumar, R. (2020), Challenges in applying machine learning models for hydrological inference: A case study for flooding events across Germany. *Water Resources Research*, 56, e2019WR025924, doi:10.1029/2019WR025924.
- Schneider, R., Koch, J., Troldborg, L., Henriksen, H. J., & Stisen, S. (2022). Machine-learning-based downscaling of modelled climate change impacts on groundwater table depth. *Hydrology and Earth System Sciences*, 26(22), 5859-5877.
- Sextstone, G. A., Clow, D. W., Stannard, D. I., & Fassnacht, S. R. (2016). Comparison of methods for quantifying surface sublimation over seasonally snow-covered terrain. *Hydrological Processes*, 30(19), 3373-3389.

- Shen, C. (2018). A transdisciplinary review of deep learning research and its relevance for water resources scientists. *Water Resources Research*, 54(11), 8558-8593.
- Shen, C., & Lawson, K. (2021). Applications of deep learning in hydrology. *Deep Learning for the Earth Sciences: A Comprehensive Approach to Remote Sensing, Climate Science, and Geosciences*, 283-297.
- Shen, Y., Liu, D., Jiang, L., Tøttrup, C., Druce, D., Yin, J., & Zhao, X. (2022), Estimating reservoir release using multi-source satellite datasets and hydrological modeling techniques. *Remote Sensing*, 14, 815, doi:10.3390/rs14040815.
- Sheridan, P., Smith, S., Brown, A., & Vosper, S. (2010). A simple height-based correction for temperature downscaling in complex terrain. *Meteorological Applications*, 17(3), 329-339.
- Shao, D., Li, H., Wang, J., Hao, X., Che, T., & Ji, W. (2022). Reconstruction of a daily gridded snow water equivalent product for the land region above 45 N based on a ridge regression machine learning approach. *Earth System Science Data*, 14(2), 795-809.
- Sicart, J. E., Pomeroy, J. W., Essery, R. L. H., & Bewley, D. (2006). Incoming longwave radiation to melting snow: observations, sensitivity and estimation in northern environments. *Hydrological processes*, 20(17), 3697-3708.
- Singh, V. P., & Woolhiser, D. A. (2002). Mathematical modeling of watershed hydrology. *Journal of hydrologic engineering*, 7(4), 270-292.
- Sinha, J., Jha, S., & Goyal, M. K. (2019). Influences of watershed characteristics on long-term annual and intra-annual water balances over India. *Journal of Hydrology*, 577, 123970.
- Sit, M., Demiray, B. Z., Xiang, Z., Ewing, G. J., Sermet, Y., & Demir, I. (2020), A comprehensive review of deep learning applications in hydrology and water resources. *Water Science & Technology*, 82, 2635-2670, doi:10.2166/wst.2020.369.
- Skøien, J. O., Blöschl, G., & Western, A. W. (2003). Characteristic space scales and timescales in hydrology. *Water Resources Research*, 39(10).
- Smyth, E. J., Raleigh, M. S., & Small, E. E. (2020). Improving SWE estimation with data assimilation: The influence of snow depth observation timing and uncertainty. *Water Resources Research*, 56(5), e2019WR026853.
- Snowdon, A. P., & Craig, J. R. (2016). Effective groundwater–surface water exchange at watershed scales. *Hydrological Processes*, 30(12), 1849-1861.
- Stigter, E. E., Litt, M., Steiner, J. F., Bonekamp, P. N., Shea, J. M., Bierkens, M. F., & Immerzeel, W. W. (2018). The importance of snow sublimation on a Himalayan glacier. *Frontiers in Earth Science*, 6, 108.
- Strasser, U., Bernhardt, M., Weber, M., Liston, G. E., & Mauser, W. (2008), Is snow sublimation important in the alpine water balance? *The Cryosphere*, 2, 53-66, doi:10.5194/tc-2-53-2008.
- Strobl, C., & Zeileis, A. (2008), Danger: High power!—exploring the statistical properties of a test for random forest variable importance. doi:10.5282/ubm/epub.2111.
- Sturm, M., & Wagner, A. M. (2010). Using repeated patterns in snow distribution modeling: An Arctic example. *Water Resources Research*, 46(12).
- Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. In *International conference on machine learning* (pp. 3319-3328). PMLR.

- Tanty, R., & Desmukh, T. S. (2015). Application of artificial neural network in hydrology—A review. *Int. J. Eng. Technol. Res*, 4(6), 184-188.
- Tartakovsky, A. M., Marrero, C. O., Perdikaris, P., Tartakovsky, G. D., & Barajas-Solano, D. (2020). Physics-informed deep neural networks for learning parameters and constitutive relationships in subsurface flow problems. *Water Resources Research*, 56(5), e2019WR026731.
- Thackeray, C. W., Derksen, C., Fletcher, C. G., & Hall, A. (2019). Snow and climate: Feedbacks, drivers, and indices of change. *Current Climate Change Reports*, 5, 322-333.
- Thapa, P., Martin, Y. E., & Johnson, E. A. (2017). Quantification of controls on regional rockfall activity and talus deposition, Kananaskis, Canadian Rockies. *Geomorphology*, 299, 107-123.
- Thornton, J. M., Brauchli, T., Mariethoz, G., & Brunner, P. (2021). Efficient multi-objective calibration and uncertainty analysis of distributed snow simulations in rugged alpine terrain. *Journal of Hydrology*, 598, 126241.
- Thornton, J. M., Therrien, R., Mariethoz, G., Linde, N., & Brunner, P. (2021). Simulating fully-integrated hydrological dynamics in complex Alpine headwaters. *Earth ArXiv*. <https://doi.org/10.31223/X5RG7Q>.
- Thornton, P. E., Shrestha, R., Thornton, M., Kao, S. C., Wei, Y., & Wilson, B. E. (2021). Gridded daily weather data for North America with comprehensive uncertainty quantification. *Scientific Data*, 8, 190, doi:10.1038/s41597-021-00973-0
- Tian, S., Zhang, X., Tian, J., & Sun, Q. (2016). Random forest classification of wetland landcovers from multi-sensor data in the arid region of Xinjiang, China. *Remote Sensing*, 8(11), 954.
- Troch, P. A., Carrillo, G. A., Heidbüchel, I., Rajagopal, S., Switanek, M., Volkmann, T. H., & Yaeger, M. (2009). Dealing with landscape heterogeneity in watershed hydrology: A review of recent progress toward new hydrological theory. *Geography Compass*, 3, 375-392, doi:10.1111/j.1749-8198.2008.00186.x.
- 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.
- Trujillo, E., Ramírez, J. A., & Elder, K. J. (2007). Topographic, meteorologic, and canopy controls on the scaling characteristics of the spatial distribution of snow depth fields. *Water Resources Research*, 43(7).
- Van Rossum, & Drake, F. L. (2009). The pickle module. Python Software Foundation. Retrieved from
- Varadharajan, C., Appling, A. P., Arora, B., Christianson, D. S., Hendrix, V. C., Kumar, V., & Zwart, J. (2022). Can machine learning accelerate process understanding and decision-relevant predictions of river water quality? *Hydrological Processes*, 36, e14565, doi:10.1002/hyp.14565.
- Varhola, A., Coops, N. C., Weiler, M., & Moore, R. D. (2010). Forest canopy effects on snow accumulation and ablation: An integrative review of empirical results. *Journal of Hydrology*, 392(3-4), 219-233.
- Vionnet, V., Verville, M., Fortin, V., Brugman, M., Abrahamowicz, M., Lemay, F., ... & Milbrandt, J. A. (2022). Snow level from post-processing of atmospheric model improves snowfall estimate and snowpack prediction in mountains. *Water Resources Research*, 58(12), e2021WR031778.
- Viviroli, D., & Weingartner, R. (2004). The hydrological significance of mountains: from regional to global scale. *Hydrology and earth system sciences*, 8(6), 1017-1030.
- Vrac, M., & Naveau, P. (2007). Stochastic downscaling of precipitation: From dry events to heavy rainfalls. *Water resources research*, 43(7).

- Wagenbrenner, N. S., Forthofer, J. M., Lamb, B. K., Shannon, K. S., & Butler, B. W. (2016). Downscaling surface wind predictions from numerical weather prediction models in complex terrain with WindNinja. *Atmospheric Chemistry and Physics*, 16(8), 5229-5241.
- Wang, G., Jia, Q. S., Zhou, M., Bi, J., & Qiao, J., Abusorrah, A. (2022), Artificial neural networks for water quality soft-sensing in wastewater treatment: a review. *Artificial Intelligence Review*, 55, 565-587, doi:10.1007/s10462-021-10038-8.
- Wang, H., Tetzlaff, D., Buttle, J., Carey, S. K., Laudon, H., McNamara, J. P., ... & Soulsby, C. (2019). Climate-phenology-hydrology interactions in northern high latitudes: Assessing the value of remote sensing data in catchment ecohydrological studies. *Science of the Total Environment*, 656, 19-28.
- Wang, J., Wu, X., Wen, J., Xiao, Q., Gong, B., Ma, D., ... & Bao, Y. (2021). Upscaling in situ site-based albedo using machine learning models: Main controlling factors on results. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1-16.
- Wang, Q., Huang, J., Liu, R., Men, C., Guo, L., Miao, Y., ... & Xia, X. (2020). Sequence-based statistical downscaling and its application to hydrologic simulations based on machine learning and big data. *Journal of Hydrology*, 586, 124875.
- Wang, S., & Peng, H. (2024). Multiple spatio-temporal scale runoff forecasting and driving mechanism exploration by K-means optimized XGBoost and SHAP. *Journal of Hydrology*, 630, 130650.
- Wang, S., Peng, H., & Liang, S. (2022). Prediction of estuarine water quality using interpretable machine learning approach. *Journal of Hydrology*, 605, 127320.
- Wang, S., Peng, H., Hu, Q., & Jiang, M. (2022), Analysis of runoff generation driving factors based on hydrological model and interpretable machine learning method. *Journal of Hydrology: Regional Studies*, 42, 101139, doi:10.1016/j.ejrh.2022.101139.
- Wang, S., Zhang, H. J., Wang, T. T., & Hossain, S. (2024). Simulating runoff changes and evaluating under climate change using CMIP6 data and the optimal SWAT model: a case study. *Scientific Reports*, 14(1), 23228.
- Wang, Z., Lai, C., Chen, X., Yang, B., Zhao, S., & Bai, X. (2015). Flood hazard risk assessment model based on random forest. *Journal of Hydrology*, 527, 1130-1141.
- Wayand, N. E. (2016). Observation, Simulation, and Evaluation of Snow Dynamics in the Transitional Snow Zone (Doctoral dissertation).
- Weierbach, H., Lima, A. R., Willard, J. D., Hendrix, V. C., Christianson, D. S., Lubich, M., & Varadharajan, C. (2022). Stream temperature predictions for river basin management in the Pacific Northwest and mid-Atlantic regions using machine learning. *Water*, 14(7), 1032.
- Whitfield, P. H. (2014). Climate station analysis and fitness for purpose assessment of 3053600 Kananaskis, Alberta. *Atmosphere-Ocean*, 52(5), 363-383.
- Wigmosta, M. S., Vail, L. W., & Lettenmaier, D. P. (1994), A distributed hydrology-vegetation model for complex terrain. *Water Resources Research*, 30, 1665-1679, doi:10.1029/94WR00436.
- Wilby, R. L., & Wigley, T. M. L. (2000). Precipitation predictors for downscaling: observed and general circulation model relationships. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, 20(6), 641-661.
- Wilby, R. L., Hassan, H., & Hanaki, K. (1998). Statistical downscaling of hydrometeorological variables using general circulation model output. *Journal of Hydrology*, 205(1-2), 1-19.

- Wilks, D. S., & Wilby, R. L. (1999). The weather generation game: a review of stochastic weather models. *Progress in physical geography*, 23(3), 329-357.
- Winstral, A., Elder, K., & Davis, R. E. (2002). Spatial snow modeling of wind-redistributed snow using terrain-based parameters. *Journal of hydrometeorology*, 3(5), 524-538.
- Wood, E. F., Sivapalan, M., & Beven, K. (1986). Scale effects in infiltration and runoff production. In *Proc. of the Symposium on Conjunctive Water Use*, IAHS Publ, No. 156.
- Wunsch, A., Liesch, T., & Broda, S. (2021). Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX). *Hydrology and Earth System Sciences*, 25(3), 1671-1687.
- Wunsch, A., Liesch, T., & Broda, S. (2022). Deep learning shows declining groundwater levels in Germany until 2100 due to climate change. *Nature communications*, 13(1), 1221.
- Xu, H., Xu, C. Y., Chen, H., Zhang, Z., & Li, L. (2013). Assessing the influence of rain gauge density and distribution on hydrological model performance in a humid region of China. *Journal of Hydrology*, 505, 1-12.
- Xu, T., & Liang, F. (2021). Machine learning for hydrologic sciences: An introductory overview. *Wiley Interdisciplinary Reviews: Water*, 8(5), e1533.
- Xu, T., Guo, Z., Liu, S., He, X., Meng, Y., Xu, Z., & Song, L. (2018). Evaluating different machine learning methods for upscaling evapotranspiration from flux towers to the regional scale. *Journal of Geophysical Research: Atmospheres*, 123, 8674-8690, doi:10.1029/2018JD028447.
- Xu, Y., Lin, K., Hu, C., Wang, S., Wu, Q., Zhang, J., ... & Luo, Y. (2024). Interpretable machine learning on large samples for supporting runoff estimation in ungauged basins. *Journal of Hydrology*, 639, 131598.
- Yang, R., Wu, J., Gan, G., Guo, R., & Zhang, H. (2024). Combining Physical Hydrological Model with Explainable Machine Learning Methods to Enhance Water Balance Assessment in Glacial River Basins. *Water*, 16(24), 3699.
- Yang, S., Yang, D., Chen, J., & Zhao, B. (2019). Real-time reservoir operation using recurrent neural networks and inflow forecast from a distributed hydrological model. *Journal of Hydrology*, 579, 124229, doi:10.1016/j.jhydrol.2019.124229.
- Yang, Y., Cao, C., Pan, X., Li, X., & Zhu, X. (2017). Downscaling land surface temperature in an arid area by using multiple remote sensing indices with random forest regression. *Remote Sensing*, 9(8), 789.
- Yin, W., Zhang, G., Liu, F., Zhang, D., Zhang, X., & Chen, S. (2022). Improving the spatial resolution of GRACE-based groundwater storage estimates using a machine learning algorithm and hydrological model. *Hydrogeology Journal*, 30(3), 947-963.
- Yoon, J., & Kavvas, M. L. (2003). Probabilistic Solution to Stochastic Overland Flow Equation. *Journal of Hydrologic Engineering*, 8, 54-63, doi:10.1061/(ASCE)1084-0699(2003)8:2(54).
- Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. *Advances in neural information processing systems*, 27.
- Yu, Y., Malone, B. P., & Renzullo, L. J. (2024). Empirical Upscaling of Point-scale Soil Moisture Measurements for Spatial Evaluation of Model Simulations and Satellite Retrievals, 2404.05229.

- Yuan, Q., Shen, H., Li, T., Li, Z., Li, S., Jiang, Y., ... & Zhang, L. (2020). Deep learning in environmental remote sensing: Achievements and challenges. *Remote sensing of Environment*, 241, 111716.
- Yuval, J., & O’Gorman, P. A. (2020), Stable machine-learning parameterization of subgrid processes for climate modeling at a range of resolutions. *Nature Communications*, 11, 3295, doi:10.1038/s41467-020-17142-3.
- Zhang, Y., Chiew, F. H., Li, M., & Post, D. (2018). Predicting runoff signatures using regression and hydrological modeling approaches. *Water Resources Research*, 54(10), 7859-7878.
- Zhao, G., Pang, B., Xu, Z., Yue, J., & Tu, T. (2018b). Mapping flood susceptibility in mountainous areas on a national scale in China. *Science of the Total Environment*, 615, 1133-1142.
- Zhao, R.-J. (1992), The Xinanjiang model applied in China. *Journal of Hydrology*, 135, 371–381, doi:10.1016/0022-1694(92)90096-E.
- Zorita, E., & Von Storch, H. (1999). The analog method as a simple statistical downscaling technique: comparison with more complicated methods. *Journal of climate*, 12(8), 2474-2489.