Advancing the efficient development and deployment of hydrologic and hydraulic models for large scale and real-time applications

by

Robert Chlumsky

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Civil Engineering

Waterloo, Ontario, Canada, 2024
© Robert Chlumsky 2024

Examining Committee Membership

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

External Examiner:	Andrew Binns Associate Professor, School of Engineering University of Guelph
Supervisor(s):	James R. Craig Professor, Dept. of Civil and Environmental Engineering University of Waterloo Bryan Tolson Professor, Dept. of Civil and Environmental Engineering University of Waterloo
Internal-External Member:	Andrea Brookfield Assistant Professor, Dept. of Earth and Environmental Sciences University of Waterloo
Internal Member:	Juliane Mai Research Associate Professor, Earth and Environmental Sciences University of Waterloo
Internal Member:	John Quilty Assistant Professor, Dept. of Civil and Environmental Engineering 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

Robert Chlumsky was the sole author for Chapters 1, 2, and 6.

This thesis consists in part of three manuscripts written for publication. Exceptions to the sole authorship of material are as follows:

Research presented in Chapter 3:

This chapter presents the results of a modelling study in which the blended model structure is applied in calibration to twelve (12) catchments within the contiguous United States. The previously calibrated blended model from *Mai et al.* (2020) was used in this chapter to provide an in-depth analysis of the calibration methodology and model performances. Robert Chlumsky further implemented 108 fixed model structures derived from the *Mai et al.* (2020) model, refining the code to efficiently run multiple independent trials of those fixed structures and two additional models, performed the analysis of results, and prepared the chapter. Drs. Mai, Craig, and Tolson previously published the blended model structure deployed in this study and its preliminary calibration (*Mai et al.*, 2020), and contributed to the design of experiments and the interpretation of results as well as the revisions to the chapter.

Chlumsky, R., Mai, J., Craig, J. R., and Tolson, B. A. (2021): Simultaneous calibration of hydrologic model structure and parameters using a blended model. Water Resources Research, 57, e2020WR029229.
doi: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020WR029229

Research presented in Chapter 4:

This chapter presents a modelling study in which the blended model structure is experimented with by altering the algorithms and process options, including new blended forcing groups. This chapter presents the development of version 2 of the blended model structure. In this chapter, Robert Chlumsky designed the modelling experiment, wrote the code to efficiently carry out the experiment, analyzed the results and prepared the chapter including handling of revisions during the review process. Drs. Mai, Craig, and Tolson previously published the blended model structure deployed

in this study (*Mai et al.*, 2020), were co-authors on the preceding work by *Chlumsky et al.* (2021), and contributed to the interpretation of results and revisions of this chapter.

Chlumsky, R., Mai, J., Craig, J. R., and Tolson, B. A. (2024): Advancement of a Blended Hydrologic Model for Robust Model Performance. Journal of Hydrologic Engineering, 29 (5), 04024,033.
doi:https://ascelibrary.org/doi/10.1061/JHYEFF.HEENG-6246.

Research presented in Chapter 5:

This chapter presents a novel hydraulic modelling and mapping approach, the Geospatially Augmented Standard Step (GASS) method, which is designed for large scales and suitable for real-time deployment. The method is implemented in the software package Blackbird and benchmarked against the HAND method and HEC-RAS 1D and 2D models. In this chapter, Robert Chlumsky wrote the Blackbird software package, designed the benchmarking experiment, gathered input data and prepared all models included in the experiment, performed the analysis of results and wrote the chapter. Dr. Craig and Dr. Tolson contributed to conceptual discussions in development of the Blackbird code, and provided their review and feedback on the chapter. Brian Peng contributed to the development of C++ versions of the Blackbird code to improve computational efficiency.

Chlumsky, R., Craig, J. R., and Tolson, B. A. (2024): A reach-integrated hydraulic modelling approach for large-scale and real-time inundation mapping. Geosci. Model Dev., preprint, gmd-2024-184.

Abstract

Hydrologic and hydraulic models are tools that may be typically applied, respectively, to simulate streamflow and to determine depths and locations of flooding. While these tools are crucially important for predicting flood events, they require niche expertise and a high degree of effort to be developed and deployed effectively.

This thesis aims to streamline the level of effort required to develop and deploy quality models within the typical workflows that support the simulation of flood events. First, the selection of optimal model structures within hydrologic models is addressed. The blended hydrologic model, which allows the selection of mathematical equations to represent processes in the hydrologic cycle to occur as part of a calibration exercise, is tested and shown to provide benefit to both model performance as well as scientific utility. Secondly, the blended model is improved through an extensive empirical experiment which delivers a high performing blended version 2 model. This model achieves high performance scores across the contiguous USA without a need to adjust the model structure, which will greatly reduce the time-consuming step in practice of manually selecting the optimal model structure for a given watershed.

Finally, a novel method for hydraulic modelling and flood mapping is introduced. Improved geospatial methods are paired with a one-dimensional hydraulic model solver and then benchmarked against conventional methods. The result is shown to provide improved accuracy of flood maps while maintaining a computational runtime that is suitable for large-scale and real-time applications. Overall, it is anticipated that this research benefits the development of crucial tools for predicting and simulating flooding.

Acknowledgments

I would like to thank my PhD supervisors, James and Bryan, for treating me like a colleague from the beginning and for enticing me back to academia from the clutches of consulting. Thank you for permitting and enabling all of my wild ideas over these past five years, including starting a company.

I want to acknowledge my PhD committee, including Andrew, Julie, John, and Andrea, for taking the time to review my thesis and providing their invaluable feedback.

A big thanks is owed to the evolving members of the hydrology research team, who have enhanced the PhD experience and provided comradery over the years.

I owe an enormous thank you to my martial arts families, and especially to Randy Dauphin, for being my mentor, friend, and so much more for nearly thirty years.

I thank my parents for showing me the value of hard work and always believing in my success, even after they stopped understanding exactly what I do.

I am forever indebted to my lovely wife Lauren for supporting me in all things always, and for seeing the vision of what could be before anyone else.

Finally, I thank my kids for keeping me grounded, for being my joy, and for inspiring me every day to do better than I thought I could.

Dedication

This thesis is dedicated to Dean, Violet, and Calvin. I hope you find your passion and chase your dreams, like I did.

Table of Contents

List of Tables **xi**

List of Figures **xvii**

List of Abbreviations **xxvi**

1 Introduction **1**
1.1 Introduction . . . 1
1.2 Thesis objectives . . . 6

2 Literature Review **8**
2.1 Importance of hydrologic model structure . . . 9
2.2 Calibration of hydrological models . . . 15
2.3 Overview of rapid floodplain mapping techniques . . . 19
2.3.1 Need for improved flood mapping in Canada . . . 19
2.3.2 Flood Models . . . 21

3 Blended Model Structure Calibration **33**
3.1 Introduction . . . 35
3.2 Methods . . . 41
3.2.1 Raven Hydrologic Modelling Framework . . . 42
3.2.2 Overview of Model Structures and Calibration . . . 45
3.2.3 Case Study using MOPEX Catchments . . . 52
3.3 Results and Discussion . . . 55
3.3.1 Calibration and Validation Performance . . . 56
3.3.2 Model Selection Performance . . . 67
3.3.3 Impact on Model Calibration with Non-Random Seeding . . . 78
3.3.4 General Discussion . . . 83
3.4 Conclusions . . . 86

4 Optimizing a general hydrologic model structure **88**
4.1 Introduction . . . 90
4.2 Methods . . . 94
4.2.1 Blended hydrologic model . . . 94
4.2.2 Blended model development and performance evaluation . . . 97
4.2.3 Model overfitting . . . 105
4.2.4 Selection of a preferred blended model configuration . . . 106
4.2.5 Validation of the selected blended model . . . 108
4.3 Results and Discussion . . . 110
4.3.1 Blended model version development and performance evaluation 110
4.3.2 Model overfitting . . . 116
4.3.3 Selection of a preferred blended model configuration . . . 119
4.3.4 Validation of the selected blended model . . . 122

4.4 Conclusions . . . 123

5 Rapid Floodplain Mapping at Larger Scales 126
5.1 Introduction . . . 128
5.2 Methods . . . 133
5.2.1 Geospatially Augmented Standard Step (GASS) overview . . . 133
5.2.2 Hydraulic modelling with the standard step method . . . 137
5.2.3 Dynamic Height Above Nearest Drainage (DHAND) . . . 139
5.2.4 Reach-integrated hydraulic properties . . . 141
5.2.5 Post-processing and flood map generation . . . 146
5.2.6 Evaluation metrics . . . 147
5.2.7 Case studies . . . 148
5.3 Results . . . 157
5.3.1 Simple Verification Benchmarks . . . 157
5.3.2 Waldemar . . . 159
5.4 Discussion . . . 163
5.4.1 Inundation forecasting requirements . . . 163
5.4.2 Relative model development effort . . . 165
5.4.3 Modelling challenges and limitations . . . 166
5.4.4 Future development . . . 169
5.5 Conclusions . . . 171

6 Conclusions 173
6.1 Conclusions . . . 173
6.1.1 Limitations . . . 176
6.1.2 Future Work . . . 178

References 180

Appendix 198

List of Tables

3.1 MOPEX12 catchment characteristics. Each of the catchments are provided with an ID (1 to 12), the USGS ID of the gauge station the watershed is draining towards, the name of that gauging station, and the state this gauge is located in. All geophysical characteristics required to setup all models described in Sec. 3.2.2 in the hydrologic modeling framework Raven are specified. These are the latitude and longitude of the center of the watershed, drainage area, elevation, slope of the terrain, the percentage of forest cover FC. The main meteorologic catchments characteristics such as mean annual sum of precipitation $\overline{P}$, mean annual sum of potential evapotranspiration $\overline{PET}$, mean annual average air temperature $\overline{T}$, mean of annual total streamflow $\overline{Q}$, the aridity index AI, and the runoff ratio RR are provided. 54

7.1 Processes and process options used for the Raven setup. In total $3 \times 3 \times 2 \times 2 \times 3 \times 1 \times 1 \times 1 \times 1 \times 1 \times 1 \times 1 = 108$ models are possible. The first option of each process M_1, N_1, O_1, P_1, Q_1, R_1, S_1, T_1, U_1, V_1, W_1, and X_1 resemble the HMETS model. All other combinations are artificial models. All process options, however, are used in different hydrologic models. The model parameters active in each option are listed as well. The ranges and a description of the parameters can be found in Table 7.2. This table is adapted from that of *Mai et al.* (2020). 199

7.2 The model parameters x_i used for the Raven setup. The parameter ranges used in calibration are provided. The process option shows where the corresponding parameter is active. The Raven table and parameter name can be used to locate the parameter in the Raven setup files. The TOPSOIL is the upper soil layer while PHREATIC is the lower soil layer. The three Raven parameters FIELD_CAPACITY TOPSOIL, SNOW_SWI_MAX, and MAX_MELT_FACTOR are derived using a sampled parameter (x_{10}, x_{14}, and x_{25}) and SAT_WILT TOPSOIL, SNOW_SWI_MIN, and MIN_MELT_FACTOR, respectively, to make sure that one parameter is always larger than the other. The baseflow coefficients BASEFLOW_COEFF TOPSOIL and PHREATIC are derived from parameters x_4 and x_{11} to allow for a logarithmic sampling. The weight-generating parameters are parameters x_{35} to x_{42} with a range of $[0 \dots 1]$, and are omitted from the table. This table is adapted from (*Mai et al.*, 2020). 200

7.3 Summary information for the MOPEX catchments as well as the 24 independent catchments selected from HYSETS for validating the blended model version 2. PET and AET are model derived estimates as this data was not available for all stations, precipitation and streamflow data are derived from historical data. All hydrological metrics are determined for the 1970-1989 period in which the models are run in this study. 205

7.4 Process options included within each blended model configuration. Model configurations 25-28 & 31-32 were used for testing weighting schemes and are omitted. It is noted that this table displays blended options only; other structural changes are described in Section 2.2 by model configuration, and subsequent model configurations may not be the same as or subsets of earlier configurations depending on other changes that may have been implemented beyond what is shown in this table. 206

7.5 Processes and process options used for the blended model version 2 setup (model configuration 36) in Raven. The model parameters active in each option are listed as well. The ranges and a description of the parameters can be found in Table 7.6. Note that weight-generating parameters used to determine process weights within each process group are not listed in this table. 207

7.6 The model parameters x_i used for the blended model version 2 setup (model configuration 36) in Raven. The parameter numbering is discontinuous as not all parameters across model configurations are used in model configuration 36. The parameter ranges used in calibration are provided. The process option shows where the corresponding parameter is active. The Raven table and parameter name can be used to locate the parameter in the Raven setup files. The TOPSOIL is the upper soil layer while PHREATIC is the lower soil layer. The three Raven parameters FIELD_CAPACITY TOPSOIL, SNOW_SWI_MAX, and MAX_MELT_FACTOR are derived using a sampled parameter (x_{10}, x_{14}, and x_{25}) and SAT_WILT TOPSOIL, SNOW_SWI_MIN, and MIN_MELT_FACTOR, respectively, to make sure that one parameter is always larger than the other. The baseflow coefficients BASEFLOW_COEFF TOPSOIL and PHREATIC are derived from parameters x_4 and x_{11} to allow for a logarithmic sampling. The weight-generating parameters, with a range of $[0 \dots 1]$, are omitted from the table. 208

7.7 Values from individual regressions applied to data separately within each MOPEX catchment correlating validation KGE to the number of parameters across multiple independent trials and model configurations. All p-values less than the 10% significance level or shown in **bold**. Regression was a basic linear regression of the form $Y \sim \beta_1 X + \beta_0$ and the p-values are associated with null hypothesis of slope $\beta_1 = 0$. 211

7.8 Values from individual regressions applied to data separately within each MOPEX catchment correlating the validation gap (maximum calibration KGE minus validation KGE of the same model run) to the number of parameters across multiple independent trials and model configurations. All p-values less than the 10% significance level are shown in **bold**. Regression was a basic linear regression of the form $Y \sim \beta_1 X + \beta_0$ and the p-values are associated with null hypothesis of slope $\beta_1 = 0$. 211

7.9 Results of the unpaired, two-sided Wilcoxon rank sum test in the independent validation catchments shown. A confidence level of 95% is used. The median KGE performance for both the original blended model configuration 0 and the selected blended model configuration 36 are shown, as well as the p-values for testing whether model config. 0 is different than the model config. 36. 213

7.10 Mean calibration and validation KGE performances for the original and v2 blended model versions for each of the 36 catchments utilized in this study. The differences in mean calibration and validation performances are also calculated. 214

List of Figures

2.1 The calculation of HAND values (H_i), based on the difference of each cell elevation (z_i) and the elevation of the nearest channel drainage elevation (z_c). Figure is adapted to include additional annotations from *Nobre et al.* (2011). 29

3.1 Map of the twelve MOPEX12 catchments located in the southeastern part of the United States. The numbering of the catchments is consistent with Table 3.1. 53

3.2 Model rank calculated from the maximum calibration NSE score for each catchment and model structure with a calibration budget of 10 000 model runs. The x-axis is the model structure rank, ranging from 1 to 111. The y-axis is the maximum calibration NSE score for each model structure. Highlighted bars indicate the position of the simple model average (SMA) (red), weighted model average (WMA) (green), blended (blue), and HMETS (violet) models. The maximum scores in the calibration period and validation period are both shown for each model structure. 58

3.3 The average model structure rank across all 240 replicates (12 catchments × 20 trials). The statistic is derived for the calibration (A) and validation (B) periods with the budget of 10 000 model runs. All 111 model structures are included in these plots while the simple model average (SMA) (red), weighted model average (WMA) (green), blended (blue), and HMETS (violet) model structures are highlighted with colored markers. 61

3.4 Model performance in calibration NSE score on the y-axis and density on the x-axis (1-109), plotted for catchments 1, 8, and 11, and faceted horizontally by the five groups of process algorithms in this study. Density axis ranges vary by plot but labels have been removed for simplicity. In each density curve the colour corresponds to a choice of process algorithm within the process group as 1, 2, or 3. Model structures 1 through 108 are fixed model structures. The horizontal grey lines indicate the NSE performance of the blended model trials. The averaged model is omitted from this analysis. Results are shown for a calibration budget of 10 000 model runs. 69

3.5 Blended model process weights from 0 to 1 on the y-axis, and numbered options for each blended hydrologic process on the x-axis, faceted by the hydrologic process group. Each box and whisker plot displays the spread of the given process weight within the process group, where horizontal lines represent the 25^{th}, 50^{th}, and 75^{th} percentiles, and outliers are shown as points. For each catchment, the data blended model result is filtered to include only information from independent trials where the calibration NSE performance was within 0.05 inclusive of the top performing independent trial. Results are shown for a calibration budget of 10 000 model runs, and only catchments 1, 8, and 11 of the MOPEX12 catchments. 73

3.6 Calibration and validation NSE scores plotted for catchments 1, 8, and 11, for the blended, blended$_{\text{HMETS}}$, and HMETS models. Column 1 presents the calibration NSE scores, and column 2 presents the validation NSE scores. Catchments results are presented by row, with catchment 1 information in row 1, catchment 8 in row 2, and catchment 11 in row 3. The blended model is shown in red, the blended$_{\text{HMETS}}$ is shown in green, and the HMETS model is shown in blue. 80

4.1 Map of the 12 MOPEX catchments and 24 additional independent catchments located in the continental United States. Captions indicate the state abbreviation and catchment index (i.e., 1 to 36). 109

4.2 Model performance in calibration and validation by model configuration. The calibration KGE is the maximum KGE across 20 independent trials, and validation KGE is the associated KGE in the validation period of that maximum trial, both averaged across 12 MOPEX catchments. . . 113

4.3 Calibration performance box plots, generated from 20 independent trials of the calibration for each model configuration and MOPEX catchment, shown for select model configurations. The orange diamonds indicate the validation performance of trial with the maximum calibration performance within the 20 independent trials (when it falls within plotting bounds). 114

4.4 Validation KGE on the y-axis (obtained from the maximum calibration KGE trial) plotted against the number of calibrated model parameters on the x-axis for 31 model configurations and 12 MOPEX catchments. A regression line for this relationship is shown in blue as a nearly flat relationship (slope of $3.2E-4$ with a corresponding p-value of 0.757 for a null hypothesis of slope $\beta_1 = 0$), with a 95% confidence interval shown in transparent grey. 117

4.5 A) Model configurations plotted based on the calibration consistency on the y-axis against the validation gap on the x-axis, measured as the maximum KGE calibration performance minus its associated validation performance, averaged across 12 MOEPX catchments. The black dashed line shows the Pareto front of non-dominated model configurations (from left to right, 36, 37, 34, 29, and 30). B) The mean of the maximum calibration KGE performance plotted against the mean validation KGE performance, averaged across 12 MOPEX catchments. Pareto points are plotted with larger shapes than other data points. 121

5.1 a) Schematic of the reach-integrated flooded area, b) plan view of the inundated area, and c) schematic of the inundated reach with interpolation of depth from streamnodes j and $j+1$. The water surface elevation is held constant along the red line in c), which is the vertical projection of the drainage path ending in the nearest drainage point. 137

5.2 HAND and DHAND layers are applied for fixed depths to compute flood extents adjacent to a river bend. A) HAND layer with shallower depth of 4.5m shows flooding in two hydraulically disconnected areas on the inside of a berm. B) Same HAND layer from plot A with a greater depth of 6.5m shows greater flooding inside the berm area as it is overtopped, and also shows a new disconnected and inundated area appears just southwest of the original area. C) A DHAND layer for the specific depth of 4.5m shows no flooding on the other side of the high berm. D) The DHAND layer for a depth of 6.5m now allows flooding on the other side of the berm as it is overtopped, and does not generate falsely inundated areas as with the HAND layer in plot B. 155

5.3 Location of the case study in the Town of Waldemar. The extents of the HEC-RAS 2D model are shown in the orange outline. Imagery obtained from Google 2020 imagery. 156

5.4 Results of the SB06 benchmark test, summarized with a profile plot for both Blackbird and HEC-RAS generated results on the left and the residual depth errors shown on the right. Errors are bounded by ±0.2m and are randomly distributed around a mean of zero. 158

5.5 Results for the three model runs evaluated against the HEC-RAS 2D model under three flow scenarios. All raster cells within a defined evaluation area were evaluated, and cells that were not consistently dry in all four models (including the 2D model) were used in the calculation of metrics. Metrics used are the mean absolute error (MAE) and normalized sum of squared errors (NSSE). Plots A and B show uncalibrated model results, and plots C and D show calibrated model results. 160

5.6 Calibrated model results compared against 2D model results, showing the results for (A) the HEC-RAS 2D model depths, (B) the HAND-Manning model classified differences, (C) classified differences for the HEC-RAS 1D model, and (D) classified differences for the Blackbird model output. 162

7.1 Observed hydrographs in each catchment by period, including the two year warm up period from January 1970 to December 1971 (grey lines), the twelve year calibration period from January 1972 to December 1983 (red lines), and the six year validation period from January 1984 to December 1989 (green lines). The hydrograph is provided for all twelve MOPEX catchments. 202

7.2 Model performance in calibration NSE score on the y-axis and density on the x-axis (1-109), plotted for catchments 1, 8, and 11, and faceted horizontally by the five groups of process algorithms in this study. Density axis ranges vary by plot but labels have been removed for simplicity. In each density curve the colour corresponds to a choice of process algorithm within the process group as 1, 2, or 3. Model structures 1 through 108 are fixed model structures. The horizontal grey lines indicate the NSE performance of the blended model trials. The averaged models are omitted from this analysis. Results are shown for a calibration budget of $10k$ model runs, and all twelve MOPEX catchments. 203

7.3 Blended model process weights from 0 to 1 on the y-axis, and numbered options for each blended hydrologic process on the x-axis, faceted by the hydrologic process group. Each box and whisker plot displays the spread of the given process weight within the process group, where horizontal lines represent the 25^{th}, 50^{th}, and 75^{th} percentiles, and outliers are shown as points. For each catchment, the data blended model result is filtered to include only information from independent trials where the calibration NSE performance was within 0.05 inclusive of the top performing independent trial. Results are shown for a calibration budget of $10k$ model runs, and all twelve MOPEX catchments. 204

7.4 Calibration performance distributions, generated from 20 independent trials of the calibration for model configurations 0 (original) and 36 (version 2) for 24 independent catchments. 212

List of Abbreviations

0D - zero-dimensional (zero-order)

1D - one-dimensional

2D - two-dimensional

3D - three-dimensional

BMSC - Blended Model Structure Calibration

CAMELS - Catchment Attributes and Meteorology for Large-sample Studies

DDS - Dynamically Dimensioned Search

DEM - digital elevation model

DHAND - Dynamic Height Above Nearest Drainage

GASS - Geospatially Augmented Standard Step

GRCA - Grand River Conservation Authority **HAND** - Height Above Nearest Drainage

HEC-RAS - Hydrologic Engineering Center River Analysis System

HMETS - Hydrological Model of École de technologie supérieure

HRU - Hydrologic Response Unit

HYSETS - Hydrometeorological Sandbox—École de technologie supérieure

MOPEX - Model Parameter Estimation Experiment

KGE - Kling Gupta Efficiency

LAI - leaf area index

MAE - mean absolute error

MCC - Matthew's Correlation Coefficient

NSE - Nash Sutcliffe Efficiency

NSSE - normalized sum of squared errors

PET - potential evapotranspiration

POTMELT - potential melt

SMA - simple model average

SWOOP2020 - Southwestern Ontario Orthophotography Project 2020

WMA - weighted model average

"Discipline finishes what motivation began."

$\sim$ Anonymous

Chapter 1

Introduction

1.1 Introduction

Floods are the most common and costly natural disasters in Canada with respect to property damage, and the most common type of flooding occurs when the flow of water in a river exceeds the channel capacity and spills over (*Bhupsingh et al.*, 2022; *Public Safety Canada*, 2024). It is expected that the impacts of flooding will increase into the future in terms of both vulnerability and hazards. Hydrologic and hydraulic models are fundamental tools available to water resources managers. Hydrologic models simulate the hydrologic cycle and are used in predicting streamflow at specified locations, while hydraulic models may be used to determine water levels and inundation extents as a function of streamflow. Whether used in a predictive or scientific capacity, these models are an abstraction of the complex natural system being simulated, and neces-

sarily simplify the representation of hydrological and hydraulic processes occurring in a watershed.

This thesis focuses on advances that will facilitate the ease of deployment of quality hydrologic and hydraulic models, with the ultimate aim of improving the overall accuracy and availability of streamflow predictions and flood simulations. Typically, these two types of models are loosely coupled where hydrologic models are used to determine streamflows at locations within the watershed, and hydraulic models are used to determine water levels and flood extents based on the flows provided from hydrologic models. With respect to hydrologic models, the focus here is on model improvement via optimization of model structure, which is defined herein as a combination of (a) the number and type of water storage units ('stores') represented in the model and (b) the collection of process equations used to quantify the hydrologic fluxes between stores. In addition to determining the optimal model structure, the values associated with process equations in the model, referred to as model 'parameters', must usually be calibrated in order to reconcile model output with historical observations. The literature provides methods for optimally estimating these parameter values through calibration (*Duan et al.*, 1992; *Tolson and Shoemaker*, 2007), for determining the sensitivity of parameters (*Razavi and Gupta*, 2015), and for assessing the ability to determine these parameters with identifiability analysis (*Wagener et al.*, 2003; *Guillaume et al.*, 2019). However, despite the knowledge that the structural representation of a model is a key and substantial component of model uncertainty (*Butts et al.*, 2004; *Gupta et al.*, 2012),

these studies generally assume a fixed model structure, and there have been very few attempts to optimize, assess the sensitivity, or estimate the identifiability of model structure. A recent study suggests that model structure is frequently selected solely based on familiarity with a particular modelling software rather than the merits of the structure for the given application (*Addor and Melsen*, 2019). This inability to scrutinize model structure can lead to large and unknown errors in important hydrological modelling applications, such as flood forecasting and the safeguarding of life and property, with limited ability to assess or reduce these errors.

In more recent literature, the first explicit calibration of model structure and parameters simultaneously in hydrologic modelling was made through the use of mixed-integer optimization by *Spieler et al.* (2020), referred to as Automatic Model Structure Identification (AMSI). This breakthrough was in part possible due to the emergence of flexible modelling frameworks, which support studies of model structure by allowing a modeller to easily change the model structure without redeveloping the entire numerical model. Following this was the introduction of the 'blended' model structure (*Mai et al.*, 2020), which generalized the structural control in the AMSI approach to adjust model structure smoothly through the use of continuous weights for individual process equations, rather than integer weights. The optimal model structure in a blended model is determined through a calibration exercise to historical observations, which traditionally involves only parameters that do not modify model structure, but in the blended model these are calibrated jointly. The use of the blended modelling

approach as a tool to identify optimal model structures, and deploy blended models as an adjustable model structure for more general use in hydrology, is explored and advanced within this thesis. Following the model calibration and other analyses to inform the model reliability, the modeller would seek to use the model for its intended purpose, which is frequently the simulation or forecasting of streamflow at one or more locations. In an application where river depths or flood extents are desired, streamflow values are passed as either peak values or complete time-series as inputs to hydraulic models.

With respect to hydraulic models for floodplain mapping, existing process-based methods can be classified into three categories based on their complexity: two-dimensional (2D) hydraulic models, one-dimensional (1D) hydraulic models, and zero-order geospatial approaches. The most complex is the 2D hydraulic modelling approach, which solve the Saint Venant equations in horizontal directions and integrate depth vertically. 1D models are the most commonly applied type of hydraulic model, and consist of a series of cross-sections to represent the landscape at specific points. The depth at each cross-section location is determined by the hydraulic model, and then spatially extrapolated to develop a 2D inundation map. Zero-order models do not simulate mass or momentum conservation, but rather rely on geospatial analysis of the terrain and linking this geospatial analysis with simple equations such as Manning's equation to rapidly produce flood maps. An increasingly popular example of this is the Height Above Nearest Drainage (HAND) model (*Nobre et al.*, 2011). Each of these methods

have their limitations when it comes to applications in flood mapping, which can be summarized as follows:

- 2D models are considered the most accurate when applied correctly but have prohibitive computational runtimes, particularly for larger-scale or real-time applications
- 1D models are considered relatively accurate and have low computational runtimes, but can fail under some challenging hydraulic conditions when assumptions are violated, including more complex terrains and in the interpolation of results to generate flood maps
- zero-order geospatial methods are generally easy to develop and fast computationally, but have limited accuracy or ability to represent more complex hydraulic conditions.

The novel approach developed in this thesis is to combine the best aspects of existing 1D hydraulic models and geospatial methods into a single approach, which maintains a fast computational runtime and accuracy, while also reducing the development effort required to configure models relative to existing modelling approaches. This method is called the Geospatially Augmented Standard Step (GASS) method, which is implemented in a software package named Blackbird and benchmarked in this thesis.

1.2 Thesis objectives

The overall goal of this research is to provide improved tools to both a) develop and calibrate hydrologic models effectively 'out-of-the-box' by addressing model structure, and b) efficiently develop hydraulic models to support floodplain mapping, particularly at large scales and real-time forecasting applications. By addressing both the estimation of streamflow through hydrologic models and the mapping of flood extents through hydraulic models, the complete workflow typically required to simulate floods through numerical models is improved upon by this thesis. Although water resources management has benefited from advancements in computing and model-building tools, modellers still spend much of their time in activities required to develop high quality hydrologic and hydraulic models rather than using them in forecasting or other management applications. This thesis aims, in part, to reduce the effort required by modellers to construct high quality hydrologic and hydraulic models.

The objectives of this thesis are:

1. To test the use of the blended hydrologic model in the context of model calibration, and demonstrate its potential to improve both overall model performance and the identifiability of model structure (Chapter 3);

2. To develop, apply, and assess an updated blended model configuration with improved performance when applied across catchments within the contiguous USA (Chapter 4); and

3. To develop and assess a new method for efficient physically-based floodplain mapping that may be useful in large-scale and real-time applications (Chapter 5).

The following chapter (**Chapter 2**) provides a literature review on available methods for model structure calibration and floodplain mapping. The thesis objectives are addressed with Chapters 3, 4, and 5, respectively. These thesis chapters are linked to published or preprint manuscripts. The thesis conclusions are provided in Chapter 6.1, along with a discussion of limitations and directions for future work.

Chapter 2

Literature Review

This chapter describes the current state of practice in both hydrologic and hydraulic models, and provides the reader with the necessary context with which to understand the remaining thesis chapters. Section 2.1 discusses hydrologic models, including the types of hydrologic model structures used in previous studies and the importance of examining model structure. This section also discusses recent advances in calibrating model structure, which ties directly into the starting point for Chapter 3. Section 2.2 examines automatic calibration methods commonly deployed in hydrologic modelling studies, which is helpful in understanding Chapters 3 and 4 of this thesis. Finally, Section 2.3 discusses background on hydraulic modelling and the current state of hydraulic modelling approaches to flood mapping.

2.1 Importance of hydrologic model structure

Hydrologic models are used for various applications, including streamflow prediction, flood forecasting, or reservoir level forecasting, or in a scientific capacity to advance our understanding of hydrologic systems. Whether used in a predictive or scientific capacity, models are an abstraction of the complex natural system being simulated, and necessarily simplify the treatment of hydrological processes occurring in a watershed, either to facilitate computational expediency or in recognition of the large degree of uncertainty regarding how the watershed functions. In most practical cases, model calibration is required in order to reconcile model output with historical observation; conventionally, this calibration process focuses on tuning only model parameters. However, hydrologic models are typically considered to have three main components that contribute to model uncertainty (*Butts et al.*, 2004; *Beven*, 2005; *Gupta et al.*, 2012): (1) input data uncertainty, (2) model parameter uncertainty, and (3) model structure uncertainty, which includes the process equations used to simulate the hydrologic cycle.

This section focuses on the importance of hydrologic model structure, including methods for determining a preferred model structure and an overview of model structures types that have been used in the literature to date. Here, model structure is referred to as a combination of a) the number and type of stores (e.g., snow or soil) represented in the model and b) the collection of process equations used to describe the relationships between fluxes and storage in each hydrological response unit (HRU).

The experiments in this thesis focus on the selection of process equations to describe fluxes, thus examining a subset of the possible extent of model structure that may be addressed by also exploring the effect of the number of stores. The term 'model structure' can also refer to other modelling decisions such as spatial discretization, but the focus in this review is on the representation of hydrologic processes and the corresponding selection of process equations for each process.

A fixed model structure is one in which the selection of represented processes and the corresponding process equations are pre-determined, and often hardcoded into the modelling software such that it cannot be adjusted by the modeller. Historically, most hydrologic models have been designed with a fixed model structure while the input data and model parameters may vary from watershed to watershed. Fixed model structures are noted to likely provide sufficient predictive performance in some catchments, but may not be adequate in all catchments or other applications (*Hoey et al.*, 2014). The fixed model structure concept is consistent with a "one model fits all" paradigm (*Fenicia et al.*, 2011), and is exemplified in manually optimized model structures such as in the GR4J model (*Perrin et al.*, 2003). This paradigm has been repeatedly called into question (*Clark et al.*, 2011; *Fenicia et al.*, 2011; *Addor and Melsen*, 2019).

In the last couple of decades, hydrologic modelling studies have identified the need for models which allow the users to change the model structure, suited to specific needs (*Leavesley et al.*, 2002; *Clark et al.*, 2011; *Fenicia et al.*, 2011). These can include the need for modifying model structure as the understanding of a system changes,

additional data justifies a more complex model (*Butts et al.*, 2004), or to build models for specific purpose-driven applications (*Andréassian et al.*, 2009; *Coron et al.*, 2012).

The emergence of numerous flexible hydrological modelling frameworks has followed these calls in the literature (*Leavesley et al.*, 1996; *Wagener et al.*, 2002; *Orellana et al.*, 2008; *Fenicia et al.*, 2008; *Clark et al.*, 2008; *Pomeroy et al.*, 2007; *Fenicia et al.*, 2011; *Kavetski and Fenicia*, 2011; *Clark et al.*, 2015a; *Kneis*, 2015; *Knoben et al.*, 2019; *Craig et al.*, 2020). These are also called Modular Modelling Frameworks (MMFs) in some studies (e.g., *Remmers et al.*, 2020). Flexible modelling frameworks have facilitated the study of model structures in hydrology, and allowed for the evaluation of multiple model structures. For example, 79 model structures were deployed and compared using the FUSE framework in *Clark et al.* (2008). *Clark et al.* (2015b) used the SUMMA framework in order to evaluate process parameterizations for three sites in the western United States. Seven distinct model structures were evaluated using the SUPERFLEX framework for four different catchments in *Kavetski and Fenicia* (2011) in order to assess the limitations of using any given fixed model structure. The SUPERFLEX framework was also used in *Fenicia et al.* (2016) to tailor model structures for individual HRUs across multiple catchments. More recent studies in hydrology continue to deploy flexible frameworks and cite their importance in facilitating research related to model structures (*Pilz et al.*, 2020; *Remmers et al.*, 2020; *Chadalawada et al.*, 2020). Such frameworks with multiple process options are becoming standard in land-surface modelling (*Cuntz et al.*, 2016).

Although now numerous flexible hydrologic models are available, the determination of a preferred or optimal model structure (i.e. model structure 'selection') is an active area of research. As stated in the conclusions of the *Knoben et al.* (2020) study, "it remains difficult to explain the type of catchments where a model might do well with attributes that quantify the catchments' ... attributes". In all reviewed studies to date that attempt to select a preferred model structure from an ensemble of fixed model structures, such efforts have not been successful (e.g., *Clark et al.*, 2008; *Chadalawada et al.*, 2020; *Remmers et al.*, 2020). In practice, the selection of a model structure is often done prior to the calibration of the model, and this decision is not revisited due to the difficulty of rebuilding model components once a structure has been selected. In a study examining more than 1 500 academic papers, (*Addor and Melsen*, 2019) found that the legacy of preferred model usage dominated model structure selection rather than the adequacy of a given model structure, a conclusion supported by the finding that the model structure used in approximately 74% of studies examined could be predicted solely by the affiliation of the lead author.

In more recent literature, the first explicit calibration of model structure and parameters simultaneously in hydrologic modelling was made through the use of mixed-integer optimization by *Spieler et al.* (2020), referred to as Automatic Model Structure Identification (AMSI). In this study, the process equations to be used by the hydrologic model (from a finite set of options) were parameterized with integer parameters, allowing the model structure to be adjusted during calibration alongside parameter

values in a mixed-integer calibration exercise. The AMSI method was also shown to be effective in identifying high-accuracy models among a pool of candidates in *Spieler and Schütze* (2024).

One issue with this approach, however, is that the information mapping between the selected process equation and its corresponding model parameters is not utilized, leading to the optimization of all model parameters, whether or not they were active in the selected model structure. This limits the efficiency of the model calibration and might cause some optimization algorithms to converge slowly or fail (*Spieler et al.*, 2020).

One way to avoid the combinatorial optimization issue is to define model structure as a continuous, rather than a discrete, selection. This is the concept introduced with a blended model structure (*Mai et al.*, 2020), in which the total flux of a process is determined from a continuous weighted average of two or more individual process equations, rather than from a single equation. This method can be applied to represent any process in the hydrologic model (e.g., infiltration, baseflow, evapotranspiration, etc.) using blended structure. This weighted averaging allows the exploration of continuous model structure space 'between' different discrete model structures, which is not possible with the integer approach of *Spieler et al.* (2020) (nor with any previously applied methods). *Mai et al.* (2020) used the blended model structure within a modified Sobol' sensitivity analysis to assess the global sensitivity of model parameters and model structure simultaneously, which is the first instance of this being done in the

literature. The study was also able to determine the relative sensitivity of individual process options as well as process groups without the need for the traditional approach of running individual model structures separately and aggregating the sensitivity estimates (*Mai et al.*, 2020). The blended model structure is referred to extensively in this thesis, and is discussed further in Chapters 3 and 4.

Although not the focus in the subsequent chapters of this thesis, an increasingly popular method of simulating streamflow with high performance is through data-driven methods. *Nearing et al.* (2021) argues that deep-learning models are able to capture complex relationships at the catchment scale, and provide superior daily streamflow predictions in ungauged basins than process-based hydrologic models in gauged basins. This may suggest that more complex relationships between hydrologic variables and responses may exist, but are currently lacking in process-based models. In process-based modelling, it is often believed that overparamaterization of a model leads to poor validation performance. However, deep-learning models tend to have orders of magnitude more parameters and appear to overcome this assumption (e.g., see *Mai et al.*, 2022a). Current efforts in data-driven models include addressing the lack of physical representation, such as implementing mass-balance constraints within their models (*Frame et al.*, 2023). Many examples of data-driven models that have been applied successfully now exist in the literature (e.g., *Nearing et al.*, 2021; *Nevo et al.*, 2022; *Frame et al.*, 2022; *Lees et al.*, 2022; *Klotz et al.*, 2022; *Arsenault et al.*, 2023), with many showing that methods such as Long Short-Term Memory (LSTM) models

may outperform process-based models in benchmarking studies (e.g., *Mai et al.*, 2022a).

The blended model is distinct from data-driven methods in that it remains process-based, physically interpretable, and inherently respects common assumptions such as the conservation of mass and energy. The blended model may also be calibrated successfully in a single basin, unlike deep-learning models. However, there are some parallels to data-driven and deep-learning methods. For example, the blended model combines the selection of process equations with the calibration of other parameters, which may be viewed as conceptually similar to the simultaneous calibration of parameters and hyperparameters in data-driven methods. The blended model also increases the number of parameters relative to traditional models, but is shown to achieve high performance without overfitting the model. The blended model may be viewed by some as a process-based method for improving performance in hydrologic models without invoking deep-learning approaches.

2.2 Calibration of hydrological models

To inform the thesis objectives which deploy model calibration experiments, this section provides a review of commonly used global calibration algorithms in the field of hydrologic modelling. The focus here is on global rather than local optimization algorithms, since the introduction of weight-generating parameters in the blended model and the interaction of model structure and parameters will require a global search algorithm to be successful in this high dimensional parameter space.

Calibration algorithms, which can be described as a non-linear minimization of the difference between observation and simulation, can generally be categorized as gradient-based, heuristic, or metaheuristic. Gradient-based (or derivative-based) search methods use the gradient or some partial first-order derivatives of model output with respect to the model parameters, which can be estimated numerically. The application of gradient-based methods to hydrological models have generally found to be nonrobust, since these methods are designed to seek local optima, and the response surfaces of hydrologic models tend to be non-linear and discontinuous (*Qin et al.*, 2018a; *Mugunthan et al.*, 2005). Cases also arise in some models where optimization with discrete or integer variables is required (e.g., *van Meerveld and McDonnell*, 2006; *Spieler et al.*, 2020), which further exacerbates the discontinuity of the response surface. A method that deploys a gradient-based scheme based on the the Gauss-Newton algorithm with additional heuristic components is the robust Gaussian-Newton algorithm (RGN) (*Qin et al.*, 2018a). The RGN has been found to improve the efficiency of the Gauss-Newton algorithm, particularly against problems with difficult objective functions (*Qin et al.*, 2018a), and provides similar or improved reliability relative to some commonly-used heuristic functions (*Qin et al.*, 2018b).

In contrast, heuristic and meta-heuristic algorithms typically do not use derivatives in their search, and provide approximate solutions to an optimization problem. The definitions of heuristic and meta-heuristic algorithms are not widely agreed upon in the literature, including in hydrology, where they are sometimes used interchangeably

(e.g., *Haddad et al.*, 2013). A heuristic algorithm is broadly defined here as one that uses a search strategy to find increasingly optimal solutions in the search space. The term 'meta-heuristic' was introduced in *Glover* (1986), and has had slightly varying definitions throughout the literature (*Stützle*, 1998; *Gavrilas*, 2010; *Blum and Roli*, 2003; *Maier et al.*, 2019). A meta-heuristic algorithm is defined here as an approximate search algorithm that employs multiple heuristic techniques and uses the information obtained in calibration to update its search strategy, where as a heuristic algorithm does not update its search strategy based on information obtained during the calibration. Heuristic (and meta-heuristic) algorithms have generally been found to be more robust in the global optimization of hydrologic models than gradient-based algorithms (*Qin et al.*, 2018a).

In considering common global optimization algorithms applied in hydrologic modelling that have shown to perform well in a large number of studies, the two algorithms that should be discussed are the Shuffled Complex Evolution (SCE-UA) (*Duan et al.*, 1992), and the Dynamically Dimensioned Search (DDS) (*Tolson and Shoemaker*, 2007) algorithms. The SCE-UA method involves the exchange of information between multiple simplexes or 'complexes' (parameter sets) that are independent (*Duan et al.*, 1992, 1993). SCE-UA has been "proven, via many studies, to be highly reliable in locating the optimum (where one exists) on the response surfaces of typical hydrological models" (*Wagener and Gupta*, 2005). However, in the benchmark study of (*Arsenault et al.*, 2014), the SCE-UA algorithm was noted to converge quickly but also become

trapped in local optima when the parameter space is large (*Arsenault et al.*, 2014). In *Tolson and Shoemaker* (2007) and in *Arsenault et al.* (2014) the performance of DDS compared favourably to that of the the SCE-UA algorithm, overcoming some of the issues with local optima.

The DDS algorithm is a simple, stochastic, single-objective optimization algorithm, that searches parameter space based on random perturbations from the current best solution. The algorithm uses a random selection of parameters as a subset to perturb within each iteration, as well as a random perturbation value. DDS is a 'greedy' algorithm since the current best solution is always used to generate the next perturbation. In addition, unlike some algorithms with a convergence stopping criteria (e.g. SCE-UA and many gradient-based methods), the DDS algorithm requires a fixed calibration budget, which is used to inform the parameter sampling in each iteration by reducing the likelihood of sampling as the calibration progresses. The DDS algorithm has been applied in many hydrologic and water resources applications (e.g., *Shafii and Tolson*, 2015a; *Seiller et al.*, 2017; *Zink et al.*, 2018; *Becker et al.*, 2019; *Yen et al.*, 2015; *Spieler et al.*, 2020), particularly in the study of computationally expensive models (*Qin et al.*, 2018b). Variations on the DDS algorithm also exist, such as the Discrete DDS (*Tolson et al.*, 2008) and the Hybrid Discrete DDS (HD-DDS) algorithm (*Tolson et al.*, 2009) designed to solve discrete, single-objective, constrained water distribution problems. A so-called modified DDS (MDDS) was presented by *Huang et al.* (2014), which modifies the parameter selection within DDS based on additional supplied sensitivity analy-

sis ranking information for each parameter. The recently introduced Chaotic DDS algorithm (CDDS) (*Wu et al.*, 2020) employs chaos theory into several components of the DDS algorithm in order to reduce the likelihood of obtaining locally optimal solutions. The Pareto Archived DDS algorithm (PA-DDS) (*Asadzadeh and Tolson*, 2013, 2009; *Asadzadeh Esfahani*, 2012) provides a multi-objective implementation of the DDS algorithm, which generates a Pareto front of non-dominated solutions.

The DDS algorithm is the one used in the remainder of this thesis, and was also used by *Spieler et al.* (2020).

2.3 Overview of rapid floodplain mapping techniques

2.3.1 Need for improved flood mapping in Canada

The need for improved flood mapping in Canada is particularly clear. Canada is the second largest country in the world by land area, with one of the lowest population densities in the world. This leads to large swaths of the country that are not mapped for flood risk at all. However, much of the population is located within flood-exposed areas, including nearly 80% of Canadian cities (*Golnaraghi et al.*, 2020), which drives to the total residential flood risk estimate of $2.9 billion CAD per year (*Bhupsingh et al.*, 2022). The mapping approach in Canada is delegated to provinces, which is further delegated to lower tier organizations in some provinces. This leads to a decentralized and fragmented approach to updating flood maps rather than a large-scale

national effort, as is more common in countries such as the USA and parts of Europe (*Thistlethwaite et al.*, 2017). Of the maps that do exist in Canada, *Henstra et al.* (2019) found that most flood maps in municipalities (62%) are considered low quality, with a median age of 18 years.

The reasons for the quality and coverage of maps are numerous and complex, though a large factor is the technical process of developing the models and updating the studies that produce these flood maps. The hydraulic models used in these applications require an engineering study with substantial effort involved in both collecting data and developing the hydraulic models that are used to delineate flood maps. As such, once an area is mapped it is typically not revisited until local budgets permit or a specific need arises, leading to the long median age of flood maps in Canada. Recent efforts by the Government of Canada through programs, such as the Flood Hazard Identification and Mapping Program (FHIMP) (*Government of Canada*, 2023), are currently providing additional resources and funding to support activities pertaining to flood mapping, though the barriers due to current technical limitations persist.

Overall, the need for improved coverage of flood maps, improved quality of flood maps, more frequent updates to flood maps, and perhaps some sharing of information nationally is clear. Public safety would also be enhanced by real-time inundation prediction, which is exceedingly rare in practice today.

2.3.2 Flood Models

The focus in this section and throughout the thesis is on fluvial flooding, which occurs when a river or channel's capacity is exceeded and water spills over, inundating the adjacent areas. However, rainfall-driven (pluvial), coastal, and floods related to water network failures that may occur as a result of other flood mechanisms (e.g. sewer backups) are all important mechanisms for flooding to consider in a broader sense.

Here, we provide background on the two main categories of approaches to flood modelling that are widely used and are the focus of this thesis, which are hydraulic models (**Section 2.3.2**) and simplified approaches to flood mapping (**Section 2.3.2**). Other approaches that exist, such as data-driven methods and those based on satellite imagery, are also presented briefly (**Section 2.3.2**).

Hydraulic models

Hydraulic models are mathematical models that simulate fluid motion, and typically require numerical solutions that are solved computationally (*Teng et al.*, 2017). The level of complexity of hydraulic models depends on the dimensionality of the differential equations. For example, the 2D hydraulic models rely on the shallow water equations, which are obtained by depth-integrating the Navier-Stokes equations and resolving the partial differential equations based on horizontal directions (x and y) as well as time (t) (*Kreitmair*, 2020). The term shallow water refers to the depth-integration

and assumption that horizontal velocity is averaged within the water column. 3D models that do not rely on this assumption also exist, but require a much higher degree of computational cost and model development effort. While 3D models may be justified in smaller scale applications such as fluvial geomorphology (*Lane et al.*, 1999) or computational fluid dynamics modelling applications (*Monaghan*, 1994), the generally accepted wisdom is that the 2D depth-integrated models are adequate for most flood modelling applications.

The depth-integrated shallow water equations for 2D models are provided in *Teng et al.* (2017) as:

$$\frac{\partial d}{\partial t} + \frac{\partial (du)}{\partial x} + \frac{\partial (dv)}{\partial y} = 0 \tag{2.1}$$

$$\frac{\partial (du)}{\partial t} + \frac{\partial (duv)}{\partial y} + \frac{\partial}{\partial x}\left(du^2 + \frac{1}{2}gd^2\right) = 0 \tag{2.2}$$

$$\frac{\partial (dv)}{\partial t} + \frac{\partial (duv)}{\partial x} + \frac{\partial}{\partial y}\left(dv^2 + \frac{1}{2}gd^2\right) = 0 \tag{2.3}$$

where d is the water depth, u and v are the horizontal velocities in the x and y directions respectively, Q is the discharge, and g is the gravitational constant. Equation 2.1 is the conservation of mass, and equations 2.2 and 2.3 are the conservation of momentum.

In one-dimension, the equations can be further simplified by assuming that flow is in a single direction parallel to the channel, which includes the floodplain area. The information at each node is traditionally represented as a cross-section or transect

across the landscape which runs perpendicular to the direction of flow. In this case, the flow is simply in the x direction, which yields the one-dimensional Saint Venant equations, presented as follows (*Teng et al.*, 2017):

$$\frac{\partial Q}{\partial x} + \frac{\partial A}{\partial t} = 0 \tag{2.4}$$

$$\frac{1}{A}\frac{\partial Q}{\partial t} + \frac{1}{A}\frac{\partial (Q^2/A)}{\partial x} + g\frac{\partial d}{\partial x} - g(S_0 - S_f) = 0 \tag{2.5}$$

where S_f is the friction slope, S_0 is the channel bed slope, and A is the cross-sectional flow area. Equation 2.4 is the conservation of mass and equation 2.5 is the conservation of momentum.

In steady-state flow, the 1D Saint Venant equations are simplified by removing time dependence, in effect assuming that the given flow is time-invariant and applied for an 'infinite' time within the model until steady-state conditions are achieved. The numerical solution for this steady-state condition is typically solved using a standard step approach, which iteratively solves depths at each location in the model along the channel. This method has been implemented in programs such as HEC-RAS for many years (*Hydrologic Engineering Center*, 2023).

For a given streamnode j, the total energy of an open channel is defined by *Cruise et al.* (2007) as:

$$z_j + d_j + \frac{\alpha_2 \bar{v}_j^2}{2g} \tag{2.6}$$

where z is the channel invert, d is the channel depth, v is the velocity, α is the velocity coefficient, g is gravitational acceleration, and the term $\left(\alpha_2 \bar{v}_j^2/2g\right)$ is called the velocity head. The specific energy of an open channel omits the elevation term z. A fundamental property of fluid hydraulics is that the same specific energy can be achieved by one of two alternate depths due to the summation of the depth and velocity head terms. The threshold depth which has no alternate depth is called the critical depth (d_c), and represents the minimum energy required to transport water for a given flow and streamnode (*Cruise et al.*, 2007). A channel with depth greater than the critical depth is considered to be under a subcritical flow regime, and depth less than the critical depth under the supercritical flow regime. During flood events, channel flow is typically under a subcritical flow regime.

The standard step method relies on solving Bernoulli's equation without the pressure head term for each pair of streamnodes:

$$z_j + d_j + \frac{\alpha_2 \bar{v}_j^2}{2g} = z_{j-1} + d_{j-1} + \frac{\alpha_1 \bar{v}_{j-1}^2}{2g} + h_e \tag{2.7}$$

where h_e is the head loss between two streamnodes. The subscripts refer to the relative location of the streamnode, where j is the upstream streamnode and $j-1$ is the downstream streamnode. It is common to assume a subcritical regime, apply a normal depth boundary condition at the downstream-most node ($j = 1$), and then proceed upstream to solve each streamnode in turn. This procedure allows the model to represent the impact of downstream streamnodes on upstream ones, such as a constriction at a

downstream streamnode which propagates upstream and results in higher depths at upstream streamnodes. This type of upstream propagation of higher depths is referred to as a backwater effect.

The head loss between two streamnodes h_e is expressed as:

$$h_e = L_e \bar{S_f} + C \left(\frac{\alpha_2 v_2^2}{2g} - \frac{\alpha_1 v_1^2}{2g} \right)$$

where L_e is the effective stream length, and C is the expansion or contraction loss coefficient, depending on whether the flow area is larger or smaller than that of the downstream streamnode (A_2). The friction slope is computed at each streamnode using the relationship:

$$Q = K S_f^{1/2}$$

where K is the total conveyance of the streamnode. The average friction slope $\bar{S_f}$ may be computed in a number of ways, including a simple average between the downstream and upstream friction slopes. Additional details on the standard step method may be found in *Hydrologic Engineering Center* (2023).

Hydraulic models are the most widely used tools for flood modelling (*Teng et al.*, 2017). 2D models have become more popular in recent years both in academia and practice, partially driven by the growing availability and reduced barriers to 2D modelling software, such as the proprietary but freely available HEC-RAS software (*Hydrologic Engineering Center*, 2023). The most important recent updates in the literature on hydraulic modelling for river flooding has focused on improving the computational

efficiency of 2D models rather than improving solution quality (*Jafarzadegan et al.*, 2023). Examples include solving reduced forms of the Shallow Water Equations, code optimization and high performance computing techniques such as parallelization and GPU acceleration. Downscaling coarse 2D results to a finer resolution has also been tried (*Schumann et al.*, 2014). Another approach targets developing 1D-2D coupled models, in which the main channel is modelled using 1D cross-sections and the floodplain with 2D grid cells (*Neal et al.*, 2012; *Bates and De Roo*, 2000). The key reason for this is that the channel requires higher resolution to represent bed geometry, while the highest velocities also generally occur in channel, leading to a grid cell resolution and timestep that is limited by the channel sections. Decoupling these so that the channel is 1D allows the floodplain to adopt a coarser mesh size and larger timestep while avoiding instabilities in the numerical solution, and may be the most efficient representation of fluvial floods (*Apel et al.*, 2009; *Jafarzadegan et al.*, 2023). Both fully 2D and hybrid 1D-2D approaches have been applied successfully (*Bates*, 2022). However, 1D-2D coupling requires addressing the challenges of both 1D and 2D within the same domain and then linking flows between the domains, which presents its own challenge and substantial level of model development effort.

Prior to advances in computing power and availability of 2D software, modellers were resigned to the simpler 1D approaches, which represent the domain using cross-sections (*Bates*, 2022). The 1D method (either as unsteady or steady flow) is still the most common method behind regulatory flood mapping. This approach is now widely

established but not deployed in many research applications. Three limitations of the 1D approach noted by *Bates* (2022) include 1) the wetting of any low-lying area within the cross-section, whether or not it is connected to the main channel, 2) the neglect of information between cross-sections, including where preferential flow paths may exist but cannot be captured in the cross-section approach, and 3) the model boundaries are defined *a priori* by the modeller, which means that the flow extents and behaviours are heavily influenced by the modeller's preconceived notions rather than allowing for emergent behaviour. The first of these issues, wetting any low-lying areas regardless of a connection to the main channel, is a property of most simplified methods as well *Enriquez et al.* (2023), which are discussed in the following section.

Simplified approaches

Another group of modelling approaches has emerged over the last decade, which are referred to herein as simplified methods. The high computational cost of hydraulic models, as well as the limitations of 1D models, lead those in fields such as forecasting (where short model run times are required) to seek alternative methods. This has led to the development of these simplified methods (*Jafarzadegan et al.*, 2023). These methods do not solve the hydraulic routing equations described in the previous section, and instead make use of geospatial algorithms and simple relationships to determine inundation. Since these do not solve the differential equations of hydraulic models these are sometimes referred to as zero-order (0D) models, though it is argued by *Teng*

et al. (2017) that this term is too focused on a hydraulics perspective.

Among these methods, *Bates* (2022) divides them into those which conserve mass, and those which conserve neither mass nor momentum. The first category is often referred to as mass-spreading models, as these models distribute water on the landscape while accounting for the total volume. These methods generally preserve landscape connections due to a cell-by-cell approach in spreading water through techniques such as cellular automata (e.g., *Guidolin et al.*, 2016; *Wijaya et al.*, 2023). One of the earliest mass-spreading methods was the Rapid Flood Spreading Model (RFSM) (*Lhomme et al.*, 2009a), which was originally designed for fluvial flood overtopping and spreading, though is also the basis for a number of pluvial flood modelling approaches (e.g., *Jamali et al.*, 2018). While these methods are substantially faster than 2D hydraulic models (100-1000 times faster than 2D models in a benchmark test by *Néelz and Pender* (2010)), the cell by cell computations may still lead to a sizeable computational time.

The non-mass conserving methods include many that rely on use of a digital elevation model (DEM) layer, such as the planar or bathtub method which intersects a series of planes over a DEM (e.g., the TVD method of *Teng et al.* (2015), updated to include mass conservation in *Teng et al.* (2019)). Methods such as AutoRoute (*Follum*, 2013) use a series of cross-sections to attempt to improve mapping of flood results. One of the most popular non-conserving methods is the Height Above Nearest Drainage (HAND) model (*Rennó et al.*, 2008; *Nobre et al.*, 2011). This method computes a HAND raster from a DEM by linking each cell to its nearest drainage point in a channel, and com-

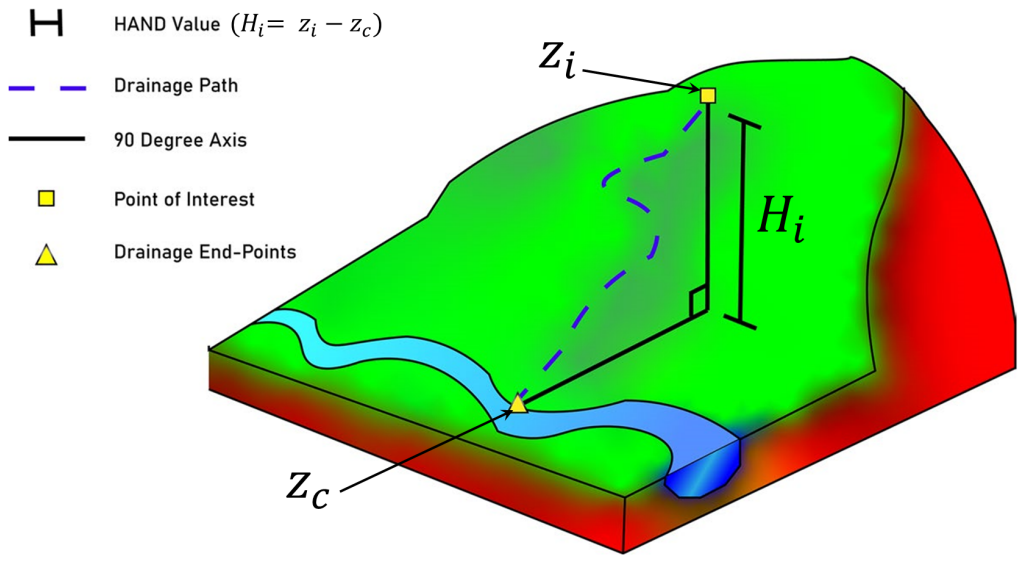


Figure 2.1 The calculation of HAND values (H_i), based on the difference of each cell elevation (z_i) and the elevation of the nearest channel drainage elevation (z_c). Figure is adapted to include additional annotations from *Nobre et al.* (2011).

puting the HAND value of the ith cell (H_i) as the differences in elevation between a cell and its channel drainage location ($H_i = z_i - z_c$, where z_i is the cell elevation and z_c is the channel elevation). This HAND raster can then be used to quickly compute flood extents for a given depth (d) by considering all cells with a HAND value less than the depth as inundated. This is shown in **Figure 2.1**, which is annotated from *Nobre et al.* (2011).

The HAND approach has been used for rapid flood inundation in many studies (e.g., *McGrath et al.*, 2018; *Chaudhuri et al.*, 2021; *Aristizabal et al.*, 2023), as well as in related applications such as estimating flood depths from satellite imagery (*Bryant et al.*, 2022). The method requires an estimation of the depth in each reach, which

is often obtained by applying Manning's equation to develop a synthetic rating curve which estimates depth from flow (e.g., *Afshari et al.*, 2018; *Scriven et al.*, 2021; *Diehl et al.*, 2021), and relies on some form of spatial averaging of hydraulic properties to estimate wetted perimeter and area in Manning's equation (*Zheng et al.*, 2018a,b). This type of approach, referred to here as the HAND-Manning method, has been shown to work reasonably well at emulating results of more sophisticated models in relatively simple cases, but struggles with more complex conditions where backwater effects influence results (*Aristizabal et al.*, 2023), since each domain or reach area the method is applied to is computed independently. The HAND method also requires a hydrologically corrected DEM, leading to the wetting of disconnected low lying areas (*Enriquez et al.*, 2023).

Other approaches

Beyond simplified geospatial methods and hydraulic models, a number of other approaches may be found in the literature, though they are not as commonly used in practice. These include a number of data-driven methods, such as the use of artificial intelligence (AI) to directly simulate flood inundation (*Shaeri Karimi et al.*, 2019), or using machine learning tools to emulate the results of complex hydraulic models with a substantially reduced computational cost at runtime, which is more often the case (e.g., *Bermúdez et al.*, 2019; *Kabir et al.*, 2020; *Chu et al.*, 2020; *Hou et al.*, 2021; *Chang et al.*, 2022; *Zhou et al.*, 2022; *Mangukiya et al.*, 2024). *Bentivoglio et al.* (2022) provides a

review of deep-learning methods for flood mapping. In a similar approach to emulation, *Fraehr et al.* (2023) developed a physics-guided surrogate model for a hydraulic model based on spatial analysis and Gaussian Process Learning. The physics-guided surrogate model was later benchmarked with other machine learning-based surrogate models in *Fraehr et al.* (2024), and found that the physics-guided surrogate models were superior. While this study found that low-fidelity physics-based models will perform better at capturing events outside of the training sets, a notion supported by the review of *Razavi et al.* (2012), it also notes that surrogate models overall have limited capability to extrapolate beyond the initial training set conditions. Machine learning techniques can also be used in data mining of online sources to determine and map active flood locations (e.g., *Lin et al.*, 2023; *Li et al.*, 2018; *Fohringer et al.*, 2015).

Satellite imagery can also be used to derive near real-time or rapid flood maps (e.g., *Tripathy and Malladi*, 2022; *Cian et al.*, 2018). These approaches are based on comparing satellite imagery during flood events to dry events, and delineating flooded areas based on the differences in imagery. Past and ongoing missions such as the Landsat, Sentinel-2, and Sentinel-1 satellites have been used in tandem towards global soil moisture and flood data collection (*Munasinghe et al.*, 2023; *Brocca et al.*, 2023). These missions have the potential to provide useful data for the calibration of inundation models, which has historically been a challenge in part due to the difficulty of obtaining reliable data sources for doing so. More recent missions, such as the Surface Water and Ocean Topography (SWOT) mission, also has the potential to enhance flood

models, and incorporating SWOT data into flood predictions is at present an active area of research (*Pietroniro et al.*, 2019; *Jet Propulsion Laboratory*, 2024).

Chapter 3

Blended Model Structure Calibration

Statement of Contribution

The following chapter is written based on the following article (*Chlumsky et al.*, 2021):

Chlumsky, R., Mai, J., Craig, J. R., and Tolson, B. A. (2021): Simultaneous calibration of hydrologic model structure and parameters using a blended model. Water Resources Research, 57, e2020WR029229.

doi: https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2020WR029229

This chapter presents the results of a modelling study in which the blended model structure is applied in calibration to twelve (12) catchments within the contiguous United States. The previously calibrated blended model from *Mai et al.* (2020) was

used in this chapter to provide an in-depth analysis of the calibration methodology and model performances. Robert Chlumsky further implemented 108 fixed model structures derived from the *Mai et al.* (2020) model, refining the code to efficiently run multiple independent trials of those fixed structures and two additional models, performed the analysis of results, and prepared the chapter. Drs. Mai, Craig, and Tolson previously published the blended model structure deployed in this study and its preliminary calibration (*Mai et al.*, 2020), and contributed to the design of experiments and the interpretation of results as well as the revisions to the chapter.

3.1 Introduction

Hydrologic models have been used for numerous applications in the last number of decades. These applications include streamflow prediction, flood forecasting, or reservoir level forecasting, or in a scientific capacity to advance our understanding of hydrologic systems. Whether used in a predictive or scientific capacity, models are an abstraction of the complex natural system being simulated, and necessarily simplify the treatment of hydrological processes occurring in a watershed, either to facilitate computational expediency or in recognition of the large degree of uncertainty regarding how the watershed functions. In most practical cases, model calibration is required in order to reconcile model output with historical observation; conventionally, this calibration process focuses on tuning only model parameters. However, hydrologic models are typically considered to have four main components that contribute to model uncertainty (e.g., *Butts et al.*, 2004; *Beven*, 2005; *Gupta et al.*, 2012): (1) input forcing data, such as precipitation and temperature, (2) model structure, which includes the algorithmic functions used to simulate the hydrologic cycle, (3) model parameters related to the selected hydrologic process algorithms (often the only degree of freedom used to tune model performance), and (4) an error term, which represents any uncertainty or deviation from reality not captured in the other three components of model uncertainty.

This paper focuses upon the simultaneous calibration of model structure and model

parameters. We refer to model structure here as a combination of a) the number and type of stores represented in the model and b) the collection of algorithms used to describe the relationships between fluxes and storage in each hydrological response unit (HRU). The emphasis here in examining model structure is specifically on the selection of process equations used to calculate fluxes. Model structure can be further extended to other modelling decisions such as spatial discretization, but this extended definition is not evaluated here. Historically, most hydrologic models have been designed with a fixed model structure while the input data and model parameters may vary from watershed to watershed. These fixed model structures were typically chosen because they (1) adequately represented the hydrologic response of one or more watersheds, (2) generally respected the physics of water flow and storage, (3) were consistent with conceptual models of the water cycle garnered from field investigation, and/or (4) convention (*Addor and Melsen*, 2019). In many cases, trial and error was used to determine final model structure (e.g., *Perrin et al.* (2003)). However, there are many such model structures that can generally meet these criteria, and as such, there exist dozens of fixed-structure models which used different algorithms to represent various components of the water cycle. The fixed model structure concept is consistent with a "one model fits all" paradigm (*Fenicia et al.*, 2011), and is exemplified in model structures such as GR4J (*Perrin et al.*, 2003). This paradigm has been repeatedly called into question (*Clark et al.*, 2011; *Fenicia et al.*, 2011; *Addor and Melsen*, 2019).

In the last couple of decades, hydrologic modelling studies have identified the need

for models which allow the users to change the model structure, suited to specific needs (*Clark et al.*, 2008; *Leavesley et al.*, 2002; *Clark et al.*, 2011; *Fenicia et al.*, 2011). Fixed model structures are noted to likely provide sufficient predictive performance in some catchments, but may not be adequate in all catchments or other applications (*Hoey et al.*, 2014). *Butts et al.* (2004) outline other reasons to advocate for flexible model structures including the ability for modellers to improve and tailor their models as their understanding of the functioning of a study site improves (e.g., *Brown and Craig* (2020)), the complexity of a model is justified, and/or as new data becomes available. An additional benefit of flexible model structures is the ability to build model structures which are purpose-driven in order to meet specific needs (*Andréassian et al.*, 2009; *Coron et al.*, 2012), or to enable scientific studies (e.g. *Remmers et al.* (2020)).

The emergence of several flexible hydrological modelling frameworks has followed these calls in the literature (*Leavesley et al.*, 1996; *Wagener et al.*, 2002; *Orellana et al.*, 2008; *Fenicia et al.*, 2008; *Clark et al.*, 2008; *Pomeroy et al.*, 2007; *Fenicia et al.*, 2011; *Kavetski and Fenicia*, 2011; *Clark et al.*, 2015a; *Knoben et al.*, 2019; *Craig et al.*, 2020). Flexible modelling frameworks have facilitated the study of model structures in hydrology, and allowed for the evaluation of multiple model structures. For example, 79 model structures were deployed and compared using the FUSE framework in *Clark et al.* (2008). *Clark et al.* (2015b) used the SUMMA framework in order to evaluate process parameterizations for three sites in the western United States. Seven distinct model structures were built using the SUPERFLEX framework for four different catch-

ments in *Kavetski and Fenicia* (2011). The SUPERFLEX framework was also used in *Fenicia et al.* (2016) to tailor model structures for individual HRUs across multiple catchments. More recent studies in hydrology continue to deploy flexible frameworks and cite their importance in facilitating research related to model structures (*Pilz et al.*, 2020; *Remmers et al.*, 2020; *Chadalawada et al.*, 2020). Such frameworks with multiple process options are standard in land-surface modelling (*Cuntz et al.*, 2016).

In the recent *Knoben et al.* (2020) study, the MARRMoT open-source framework was used to facilitate the study of 36 lumped hydrologic model structures for 559 catchments in the CAMELS data set. The model performance was found to be uncorrelated with the number of model parameters, and instead related to the choice of model parameterization (i.e. the given model structure and calibrated parameter values). While the model structure performance was found to have the strongest correlation to streamflow regime, there was little correlation found for model structure performance based on catchment characteristics, such as soil, vegetation, topographic, and geologic attributes. As stated in the conclusions of *Knoben et al.* (2020), "it remains difficult to explain the type of catchments where a model might do well with attributes that quantify the catchments' geologic, topographic, soil, and vegetation attributes". It is worthwhile to note that in all of these studies, the focus is on evaluating the performance of a pre-selected set of model structures, rather than optimally identifying the best structure. It is also noted that in all previous studies using the enumeration of fixed model structures to select an optimal structure, the selection of an optimal model

structure has not been successful (e.g., *Clark et al.* (2008); *Chadalawada et al.* (2020); *Remmers et al.* (2020)).

In more recent literature, the first explicit calibration of model structure and parameters simultaneously in hydrologic modelling was made through the use of mixed-integer optimization by *Spieler et al.* (2020), referred to as Automatic Model Structure Identification (AMSI). Here, integer option numbers indicating the hydrologic process algorithm to be used by the hydrologic model (from a finite set of options) were defined and explicitly calibrated as part of the mixed-integer calibration exercise. Along with these integers, the associated continuously valued model parameters were calibrated, identifying not only the optimal model structure but also the optimal parameterization of this model. This represents another step forward into explicitly calibrating and effectively selecting model structure. However, one limitation of the AMSI approach is that the information mapping between the selected process algorithm and its corresponding model parameters is not utilized, leading to the optimization of all model parameters, whether or not they were active in the selected model structure. This limits the efficiency of the model calibration and might cause some optimization algorithms to converge slowly or fail (*Spieler et al.*, 2020). The AMSI approach also requires a mixed integer optimization algorithm (handling both continuous and integer valued decision variables) to solve the joint calibration problem and while such algorithms exist, the vast majority of optimization algorithms employed for hydrologic model calibration are only for continuous valued decision variables.

Here, the automated simultaneous structure-parameter calibration approach of *Spieler et al.* (2020) is generalized by calculating various hydrologic fluxes using a continuous weighted average between fluxes calculated from individual process algorithms (i.e., different structures). This idea was introduced by *Mai et al.* (2020) for simultaneous model structure and parameter sensitivity analysis. For instance, rather than choosing between two discrete process algorithms for baseflow, the baseflow flux may be calculated as the weighted average of that generated from two available baseflow algorithms. This may be extended for any available process algorithm grouping with a finite number of available process algorithms. This weighted averaging allows the exploration of continuous model structure space 'between' different discrete model structures, which is not possible with the integer approach of *Spieler et al.* (2020). Lastly, this weighted averaging approach also partially addresses the issue of calibration algorithm inefficiency due to conditionally inactive parameters as noted by *Spieler et al.* (2020), whereby a large number of parameters can be inactive and thus non-influential in unused process options.

This study uses the concept of blended model structures proposed by *Mai et al.* (2020) directly within a calibration framework, enabling efficient simultaneous calibration of structure and parameters. The objectives of the study are to 1) assess the efficacy of this blended model structure calibration (BMSC) approach on a number of MOPEX catchments, and 2) compare the computational efficiency of the BMSC approach to independently calibrating a finite sample of different model structures. In

meeting these objectives, we intend also to address the following research questions:

1. Can blended model structures defining model structures using continuous parameters be used to simultaneously calibrate model parameters and model structure? Can the calibration of blended models enhance the overall performance of existing hydrologic models?

2. Can calibrated blended model structures reveal preference for one model structure over another?

3. What is the importance of seeding on the blended model calibration with a good initial model structure?

In Section 3.2 the methods, data, and case study domain will be introduced. The study results and discussion for the study objectives and research questions will be presented in Section 3.3. Finally, the study conclusions, limitations, and suggestions for future work will be identified in Section 3.4.

3.2 Methods

Section 3.2.1 will introduce the Raven Framework, the software used in this study to construct hydrologic models and present the method of blended model structure weighting. The details of the model structures used, the calibration schemes, and the modelling experiment are discussed in Section 3.2.2. Finally, MOPEX catchments case study details are provided in Section 3.2.3.

3.2.1 Raven Hydrologic Modelling Framework

The Raven hydrologic modelling framework (Raven) is an open-source, object-oriented software framework that can be used to build models from a selection of more than 100 available process algorithms (*Craig et al.*, 2020). *Mai et al.* (2020) calculates that approximately 8×10^{12} hydrologic model configurations may be setup using Raven. For this study, the benefits of using Raven include (1) the large number of available model structures, (2) the computational efficiency of the software and (3) the unique support for blended model structures, which is necessary to support the proposed calibration approach. With this blended model option, the user may specify process weights for multiple hydrologic flux algorithms, and Raven will internally use a weighted sum of the calculated fluxes in each model simulation time step. The blended approach respects the water balance, as the weighted flux is still limited by the water availability and deficits in the relevant storage compartments, and is handled the same as non-weighted fluxes in Raven.

Importantly, the process flux blending is done during a single model simulation, rather than using multiple runs with different process algorithms and combining the outputs. This blending is done within defined process groups with two or more hydrologic flux algorithms, in which every grouped algorithm affects the same flux term. Defining such process groups requires expert guidance to ensure the algorithms are appropriately grouped within their given process groups, i.e. that each process algorithm within a group moves water between the same state variables. It needs to be noted

that this process flux blending approach is fundamentally different from obtaining a multi-model simulation using the post-processing and averaging of model outputs, as the weighting is done internally within the hydrologic model on each time step for each hydrologic process. The weighting of process options within a single run in Raven has only a nominal impact on the model run time, estimated in this experiment as a 2.3% increase in the average model executable runtime relative to a non-blended model run.

The approach of blended model structures used here was introduced by *Mai et al.* (2020), and uses the exact same set of blended process groups. In this setup, Raven is applied as a lumped model with two options for P_i for the baseflow process, two options O_i for evapotranspiration, three options M_i for infiltration, three options N_i for quickflow, and three options Q_i for snow balance. Other processes are included in the model but are not blended, and are referred to as additional processes X_1. These blended process options result in an enumeration of 108 plausible fixed model structures when blending of processes is not performed. The fixed model structure generated by selecting the first process option in each blended group ($i = 1$) results in an emulation of the HMETS model (*Martel et al.*, 2017). The HMETS model is a relatively simple open-source model that can be perfectly emulated using Raven, and the HMETS process algorithms are easily grouped with other algorithms in Raven for blending, which makes the HMETS model a good candidate for inclusion as a benchmark in this study.

Within each process group, i weights (w) are required, where i is the number of

process options ($NP_{\mathcal{G}}$) within each process group $\mathcal{G}$. For example, the total flux $f_{\mathcal{G}}(t)$ of the infiltration process $\mathcal{G} = M$ within the blended model structure in a given time step t may be determined as:

$$f_M(t) = \sum_{i=1}^{NP_M} w_{Mi} f_{Mi}(t) \tag{3.1}$$

where $f_{Mi}(t)$ is the ith infiltration flux provided by infiltration process option i, w_{Mi} is the corresponding weight, and NP_M is the number of process options for infiltration (in this setup, 3). This weighting is extended to each of the five blended process groups in this configuration (infiltration M, quickflow N, evapotranspiration O, baseflow P, and snow balance Q).

This weighted representation can be extended to any model processes with multiple process algorithms available. In the Raven model files, a process group can be specified with NP algorithms per group ($NP \geq 1$), and NP process weights are provided to determine internally the overall weighted flux following Eq. 3.1. These weights w_i ($i = 1, \ldots, NP$) need to sum up to one. Therefore, the weights, now parameters of the model, cannot be sampled independently during calibration. Therefore, $NP - 1$ independent weight-generating parameters r_j ($j = 1, \ldots, NP - 1$), each between zero and one, are introduced which may be sampled during calibration (i.e., decision variables for the optimization algorithm), and are transformed within Raven into the weights w_i. The transformation approach ensures that 1) the weight-generating parameters sampled during calibration may be adjusted independently, such that the order of the weight-generating parameters is not relevant, and 2) the weights in Raven will sum

to unity. The relationship between the set of weight-generating parameters and set of weights is such that every r_1, r_2 pair uniquely generates a set of $w1$, w_2, and w_3, and any w_1, w_2, and w_3 uniquely corresponds to a single pair r_1, r_2, and r_3. The transformation approach was first introduced by *Moeini et al.* (2011) and is provided in detail in Appendix A of *Mai et al.* (2020). The blended model weights therefore add a total of eight additional model parameters (weight-generating parameters) with this approach, which are used to generate 13 weights internally in Raven. This leads to an overall number of 42 parameters of the blended model structure (versus 34 parameters collectively in the fixed model structures).

3.2.2 Overview of Model Structures and Calibration

This study compares 108 fixed model structures generated from the permutation of multiple process algorithms within five different hydrologic processes. These included: two options for baseflow, two options for evapotranspiration, three options for infiltration, three options for quickflow, and three options for snow simulation (termed "snow balance"). Additional hydrologic processes with fixed process options, such as rain-snow partitioning and percolation, are also included in each of the discrete model structures to form a complete model.

This set of structures and process algorithms are produced from the setup introduced and analysed by *Mai et al.* (2020). For completeness the process options and associated parameters including their ranges can be found in the **Appendix** (**Table 7.1**

and **Table 7.2**). Each model structure is setup within the Raven Hydrologic Modelling Framework, and the process names and algorithmic descriptions may be found in the Raven Manual (*Craig*, 2020). One model structure within the set of 108 of the discrete model structures is the HMETS model (*Martel et al.*, 2017). At least one of the five "HMETS" process algorithms is included in 100 of the 108 fixed model structures in this study.

In order to provide an additional set of benchmarks for comparison within the study, additional averaged models generated from an average of the 108 fixed hydrograph simulations were included in the study. The simple model average (SMA) hydrograph is calculated as the mean of the simulated hydrographs of all 108 fixed model structures using equal weighting. The weighted model average (WMA) hydrograph is calculated by the approach used in *Lane et al.* (2019), in which all calibrated models with a Nash-Sutcliffe Efficiency (NSE) score above 0.5 are assigned a normalized weight, and all other models are given a weight of zero. The WMA is considered to be more informed by the calibration results of the fixed model structures than the naive average performed to obtain the SMA. These models represent typical multi-model approaches in literature, (e.g., *Duan et al.*, 2007; *Velázquez et al.*, 2010; *Diks and Vrugt*, 2010; *Seiller et al.*, 2017; *Muhammad et al.*, 2018; *Darbandsari and Coulibaly*, 2019; *Lane et al.*, 2019).

Finally, the 'blended' model structure is included using the weighted average of all available process options, as described in Section 3.2.1. The model weight-generating

parameters are calibrated simultaneously alongside other model parameters during the calibration process. This averaging of process fluxes from multiple algorithms is performed during the model runtime at each timestep and whenever each process is simulated, and is is therefore fundamentally different from the 'averaged' model which averages the final simulated streamflow in post-processing. This blended model was used by *Mai et al.* (2020) in the context of sensitivity analysis, but has otherwise not been deployed for calibration prior to this study.

This blended model is the main focus of this study, specifically in determining whether the blended model structure can be successfully deployed as an alternative to enumerating fixed model structures (such as the 108 structures), and whether this blended approach may enhance the overall performance of the hydrologic models. It must be emphasized that when we compare the blended model (one calibration problem) to the 108 alternative structures, this is a scenario where the modeller is unsure about the best structure. Hence, the modeller must build 108 different models and then solve 108 calibration problems in order to select a 'best' model structure.

A total of 111 models and structures will be compared during this study:

- 108 fixed model structures, based on the enumeration of permutations from the five processes with multiple possible algorithms, one of which emulates the HMETS model;
- the blended model structure, which is the result of simultaneous calibration of structure and parameters; and

- two 'averaged' models (the SMA and WMA), which both utilize some average of the 108 fixed model structure hydrographs, representing an ensemble approach.

Each model configuration uses a lumped spatial representation and a daily time step. Identical forcing inputs, state initializations, and process algorithm implementations are provided for each model structural configuration.

The model evaluation metrics for calibration is calculated based on twelve years of simulated catchment outflows over the period from January 1, 1972 to December 31, 1983. The model simulations for calibration are performed following a two-year warm-up period from January 1, 1970 to December 31, 1971. The model validation period of six years follows the calibration period, and spans from January 1, 1984 to December 31, 1989. The calibration and validation periods are within the typical range found in the literature for similar studies, such as the two year calibration and one year validation period in *Kavetski and Fenicia* (2011), the ten year periods in other studies (*Rakovec et al.*, 2019; *Mizukami et al.*, 2017; *Kratzert et al.*, 2019a), and the 20+ year calibration and validation periods in *Spieler et al.* (2020).

The calibration was performed using the Dynamically Dimensioned Search (DDS) algorithm (*Tolson and Shoemaker*, 2007), as implemented in the Ostrich calibration toolbox (*Matott*, 2017). DDS is ideally suited for computationally expensive problems (*Tolson and Shoemaker*, 2007), and has been used in many studies to calibrate hydrological models (*Shafii and Tolson*, 2015b; *Seiller et al.*, 2017; *Zink et al.*, 2018; *Becker et al.*, 2019; *Spieler et al.*, 2020). The calibration of each of the 109 model configu-

rations that require calibration is completed with a prescribed budget as required by the DDS algorithm. Initial values in the calibration were set as random starting values within the available parameter ranges for all calibration runs. All model structures are calibrated using a DDS budget of 10 000 iterations. This is based on a common upper limit for large computation studies found in literature, such as: the 10 000 iterations with the DDS algorithm in *Tolson and Shoemaker* (2007), the 1 000 iterations budget for DDS and 10 000 iterations limit for the Shuffled Complex Evolution (SCE) algorithm in *Rakovec et al.* (2019), and the convergence with the SCE algorithm after 16 000 iterations to 18 000 iterations in *Cuntz et al.* (2015). A maximum budget of 10 000 was also used in the comparable *Spieler et al.* (2020) study. All model calibration runs were repeated twenty times with different randomly generated initial parameter sets in order to obtain a distribution of results.

The calibration budget of 10 000 does not guarantee convergence of the algorithm; DDS is not designed to converge to the precise global optimum, and it adjusts its parameter sampling according to the prescribed calibration budget. This potential for any given calibration run to not converge to the precise global optimum within the specified calibration budget allows for model structures to outperform a more complex model structure that they are a subset of. Specifically, in this experiment all fixed model structures are subsets of the blended model structure, and could be reproduced exactly with the correct set of parameters and weights in the blended model. If all model structures converged to the global optimum, then the blended

model would perform equally well or even better than all model structures in this experiment. However, with the specified calibration algorithm and budget, a fixed model structure may outperform the blended model. The fixed calibration budget for each model structure, despite the larger number of parameters with some models (such as the blended model), also increases this possibility. The use of the best calibration result out of 20 optimization trials mitigates against this possibility.

The models are calibrated regarding the NSE metric (*Nash and Sutcliffe*, 1970):

$$NSE = 1 - \frac{\sum_{t=1}^{T} (y_t - \hat{y}_t)^2}{\sum_{t=1}^{T} (y_t - \bar{y})^2} \tag{3.2}$$

where y_t is the simulated value at time t, $\hat{y}_t$ is the observed value at time t while T is the number of observations, and $\bar{y}$ is the average of the T observed data. The NSE is bounded as $[-\infty, 1]$ and is useful in interpretation of relative performance, since a value of 1 indicates a perfect fit to the observed data while negative NS efficiencies indicate a performance worse than simply using the average of the observed data.

The NSE metric, as well as other integrated metrics, has been criticized in the literature as providing a relatively weak benchmark for hydrologic performance (*Gupta et al.*, 2008), as well as providing insufficient information to discriminate hydrologic performance between model structures (e.g., *Shafii et al.*, 2017). The authors generally agree with these limitations of integrated metrics such as the NSE. Model consistency would also be improved by the use of a Pareto-based approach with multiple objectives in optimization (*Shafii and Tolson*, 2015b), ideally with observed data sources beyond streamflow (e.g., snowpack measurements). However, for the purpose of providing a

comparison for a large number of model simulations, integrated metrics based on historical streamflow records provide a necessary first-level overview of the model performances across thousands of model runs. The NSE is used in this paper to demonstrate those differences. In meeting the purpose of this study to demonstrate the simultaneous calibration of model structure and parameters, the integrated metric is deemed to be sufficient. The NSE metric was also used in the description of the MOPEX12 model performances in *Duan et al.* (2006) and as the objective function in the AMSI method of *Spieler et al.* (2020).

It is also noted that the selection of DDS as the calibration algorithm may impact the performance in this experiment. As noted by *Behrangi et al.* (2008), DDS may not be the ideal algorithm for a large number of iterations where fine-tuning of parameters is required for the 'optimal' solution. However, the strength of DDS is to find 'good' quality solutions globally (*Tolson and Shoemaker*, 2008), which is helpful in this study to help ensure the model performance across many model structures is not impacted by local optima. In any case, the purpose of this study is to compare the relative performance across model structures including the 'blended' model structure, which is achieved by the use of a consistent optimization method for all structures. For the purpose of fine-tuning parameters for further improvements in model performance, pairing the DDS algorithm with a subsequent gradient-based method, such as the Nelder-Mead algorithm (*Nelder and Mead*, 1965) may be desirable.

3.2.3 Case Study using MOPEX Catchments

The Model Parameter Estimation Experiment (MOPEX) is an international project aimed at developing *a priori* parameter estimations in hydrologic models. The MOPEX database contains 438 catchments across the continental United States. During the second and third MOPEX workshop twelve of those catchments were selected as a benchmark set. This set is referred to in this study as the 'MOPEX12'. Details of the second and third MOPEX workshops may be found in *Duan et al.* (2006).

These twelve catchments cover the southeastern portion of the United States and were selected since they span a wide range of hydrologic characteristics (*Duan et al.*, 2006), making them a suitable set of catchments for testing various model structures on a range of hydrologic conditions. The locations of these catchments are provided in Figure 3.1. The catchment characteristics of the MOPEX12 are provided in Table 3.1, calculated from the available 56 years of forcing data from 1948 to 2003. Besides major basin characteristics such as mean annual temperature, precipitation and evapotranspiration, the table specifies the geophysical properties required to set up the models in Raven.

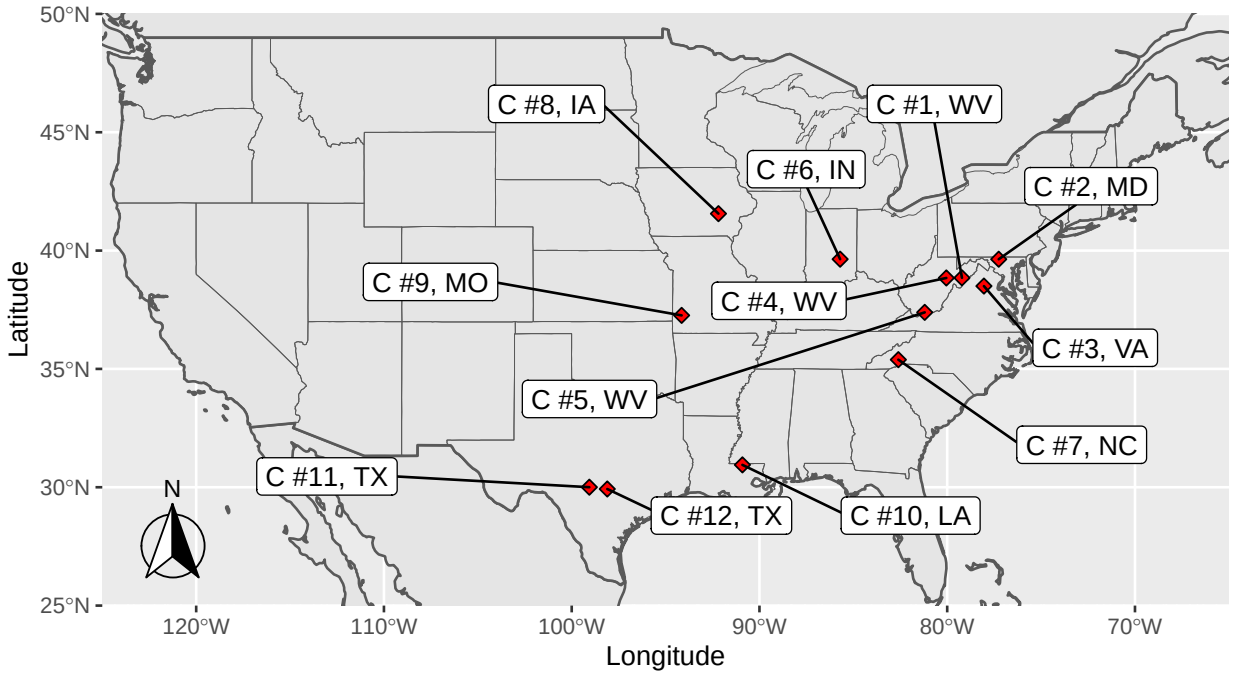


Figure 3.1 Map of the twelve MOPEX12 catchments located in the southeastern part of the United States. The numbering of the catchments is consistent with Table 3.1.

Table 3.1 MOPEX12 catchment characteristics. Each of the catchments are provided with an ID (1 to 12), the USGS ID of the gauge station the watershed is draining towards, the name of that gauging station, and the state this gauge is located in. All geophysical characteristics required to setup all models described in Sec. 3.2.2 in the hydrologic modeling framework Raven are specified. These are the latitude and longitude of the center of the watershed, drainage area, elevation, slope of the terrain, the percentage of forest cover FC. The main meteorologic catchments characteristics such as mean annual sum of precipitation $\overline{P}$, mean annual sum of potential evapotranspiration $\overline{PET}$, mean annual average air temperature $\overline{T}$, mean of annual total streamflow $\overline{Q}$, the aridity index AI, and the runoff ratio RR are provided.

ID	USGS Gauge ID	Basin Name	Lat.	Lon.	Area [km^2]	Elev. [m]	Slope [°]	FC [%]	$\overline{P}$ [mm]	$\overline{PET}$ [mm]	$\overline{T}$ [°C]	$\overline{Q}$ [m^3/s]	AI [-]	RR [-]
1	01608500	South Branch Potomac River, WV	38.85	-79.21	3782	661.4	11.0	8.1	935	762	10.7	1414	1.23	0.40
2	01643000	Monocacy River, MD	39.63	-77.26	2115	194.6	3.1	0.7	1059	898	11.5	847	1.18	0.38
3	01668000	Rappahannock River, VA	38.50	-78.04	4129	207.3	4.2	5.5	1078	921	12.5	1481	1.17	0.33
4	03054500	Tygart Valley River, WV	38.84	-80.05	2366	718.7	8.6	8.6	1318	711	9.8	1755	1.85	0.56
5	03179000	Bluestone River, WV	37.39	-81.20	1023	793.2	8.3	4.1	984	741	11.0	418	1.33	0.42
6	03364000	East Fork White River, IN	39.64	-85.70	4419	268.8	0.6	0.0	1032	856	11.0	1681	1.21	0.37
7	03451500	French Broad River, NC	35.39	-82.60	2446	829.8	8.8	3.1	1510	820	12.5	1831	1.84	0.50
8	05455500	English River, IA	41.56	-92.18	1486	253.5	1.6	4.0	867	995	9.8	357	0.87	0.28
9	07186000	Spring River, MO	37.26	-94.15	3013	331.5	0.8	0.0	1079	1095	13.7	856	0.99	0.26
10	07378500	Amite River, LA	30.95	-90.91	3314	82.4	1.0	7.2	1575	1074	18.9	1831	1.47	0.35
11	08167500	Guadalupe River, TX	30.00	-99.06	3404	544.4	2.7	0.0	772	1529	18.2	359	0.51	0.14
12	08172000	San Marcos River, TX	29.93	-98.10	2169	298.8	1.8	0.3	844	1449	19.5	361	0.58	0.20

The MOPEX12 catchments were selected for use in this study in order to evaluate the performance of each model structure in different catchments and geographic locations. Each model structure was applied using the same set of respective parameters, calibration and validation periods, etc., and models were calibrated separately (i.e. with different calibrated parameter values) for each of the twelve catchments. The same set of MOPEX12 catchments have been used in other studies, such as *Evin et al.* (2014) and *Cuntz et al.* (2016), and additional catchment properties may be found therein. The MOPEX catchments have also been used to support numerous other studies (*Clark et al.*, 2008; *Kavetski and Clark*, 2010; *Ye et al.*, 2014). The continued use of MOPEX catchments in scientific studies contributes to best practices with respect to open data and reproducability of results.

3.3 Results and Discussion

We present the results of the model performances across 111 model structures, including the 108 fixed, one blended, and two averaged model structures. First, the overall model performances are presented and discussed (Section 3.3.1). These address the primary study objectives of determining whether blended model structures can be used to simultaneously calibrate model parameters and structures, and how this compares to the calibration of fixed model structures. Second, the modelling results are explored to provide insight on dominant processes and model structures across spatially and hydrologically variant watersheds. The role of the blended model in this analysis is

discussed in Section 3.3.2. Finally, the impact of non-random initial values on the blended model is explored in Section 3.3.3).

3.3.1 Calibration and Validation Performance

A total of 109 model structures, including the 108 fixed and one blended model structure, are calibrated with the 10 000 calibration budget for the twelve MOPEX catchments, with 20 independent trials for each run. This results in 240 (12×20) calibration exercises for each of the 109 model structures. The hydrographs of the averaged models are derived as an average of the hydrographs of 108 optimal fixed model structures. The primary objective of this experiment is to determine whether the blended model with continuous structural-weighting parameters can be successfully deployed and used to calibrate model parameters and structure. The relative performance of the blended model structure to that of the fixed model structures and averaged structures under the same calibration conditions is also discussed.

In total, twenty independent trials are performed in order to account for the random component of the calibration procedures. The twenty results hence lead to a distribution of calibrated model results rather than a single set of results. The calibrated parameters sets (including the parameters used to describe the model structure in the blended model) are used to derive model results for the calibration (1972 to 1983) and validation (1984 to 1989) period. Recognizing that the DDS algorithm is designed to achieve a "good" global performance (a "good" solution as per the definition in *Tolson*

and Shoemaker (2008)), a common calibration strategy when deploying DDS is to use the best performing optimization trial from a number of independent trials. This strategy is deployed here by taking the optimization trial with the maximum calibration score in the reporting of results. The maximum calibration scores of the 20 trials for the twelve catchments and their corresponding validation scores are shown in **Figure 3.2**.

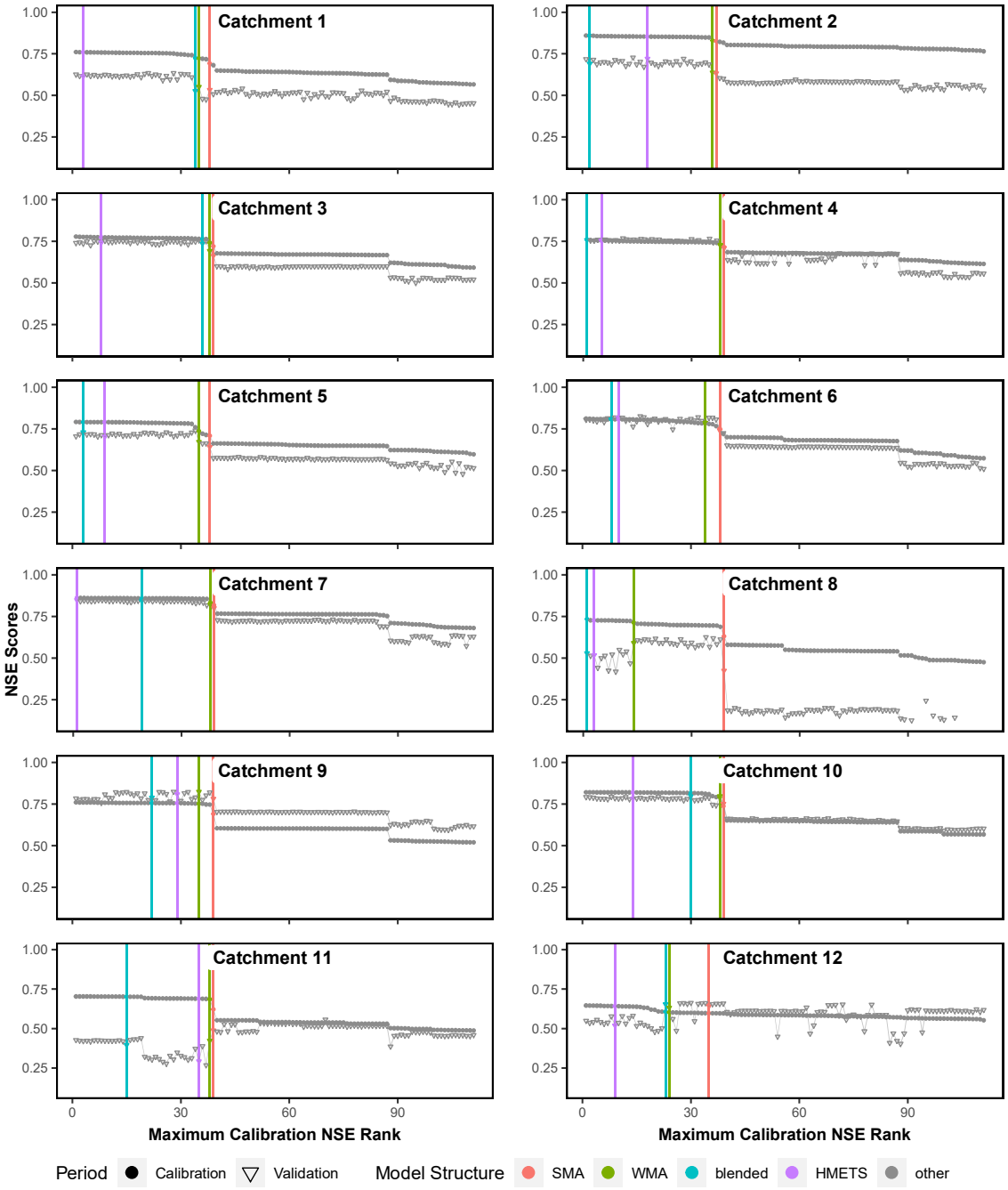


Figure 3.2 Model rank calculated from the maximum calibration NSE score for each catchment and model structure with a calibration budget of 10 000 model runs. The x-axis is the model structure rank, ranging from 1 to 111. The y-axis is the maximum calibration NSE score for each model structure. Highlighted bars indicate the position of the simple model average (SMA) (red), weighted model average (WMA) (green), blended (blue), and HMETS (violet) models. The maximum scores in the calibration period and validation period are both shown for each model structure.

In **Figure 3.2** the blended model is within the top quartile for nearly all twelve MOPEX catchments, and ranks as the top model structure in catchments 4 and 8. The blended model achieves a higher maximum calibration NSE than both averaged models in all catchments, and achieves a larger calibration NSE than the HMETS model in seven of the twelve catchments. The HMETS model tends to rank within the top quartile for most of the catchments, with the exception of catchment 11, and is the top ranked model in catchment 7. The HMETS model also achieves a higher maximum calibration NSE than the averaged model in all catchments. The SMA model tends to lag in performance rank in comparison to the blended and HMETS model structures, with a consistent rank near position 35. The calibration-informed WMA model consistently outperforms the SMA model in calibration for all 12 catchments.

Another way in which the results were analysed is by evaluating the average rank of each of the 111 models over all 240 replicates (12 catchments $\times$ 20 trials), which uses the calibration performance from all independent trials rather than the maximum of each set. Each model is assigned a rank (1 to 111) based on the NSE for the period analyzed (calibration or validation) for each trial. The results of the average rank across those 240 experiments are derived for both calibration and validation period (**Figure 3.3**). Every point in each subpanel represents one of the 111 model structures considered. The y-axis is the raw average rank (across 240 replicates) while the x-axis is simply the ranks of these raw average ranks. For example, considering panel (A), the blended model has the best average rank across replicates (6.1 on the y-axis) and is thus ordered

as rank 1 on the x-axis. In comparison, panel (A) shows the SMA model has an average rank of 24.4 on the y-axis and that average rank is 13^{th} best, and thus ranked 13^{th} on the x-axis.

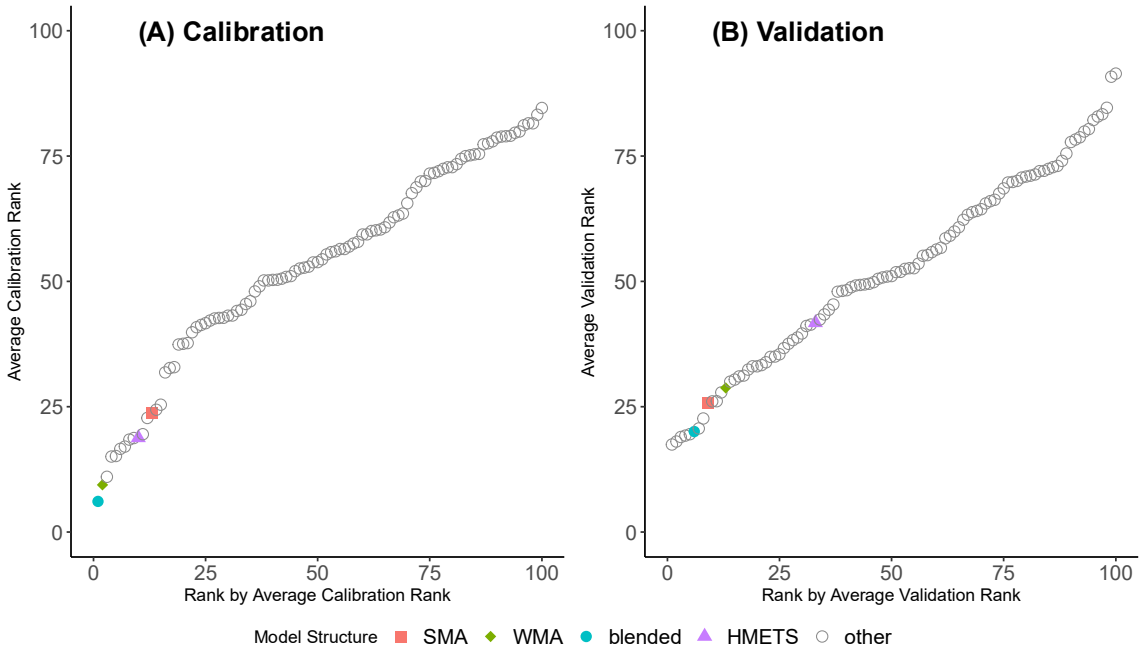


Figure 3.3 The average model structure rank across all 240 replicates (12 catchments × 20 trials). The statistic is derived for the calibration (A) and validation (B) periods with the budget of 10 000 model runs. All 111 model structures are included in these plots while the simple model average (SMA) (red), weighted model average (WMA) (green), blended (blue), and HMETS (violet) model structures are highlighted with colored markers.

In **Figure 3.3**, the blended model is the top ranked model structure based on the calibration NSE performance, when averaged over all catchments and trials; the WMA model is the second ranked model structure. In calibration, the blended, HMETS, and averaged models all rank within the top 25 positions by average calibration rank. However, in validation, the HMETS model has a reduced performance relative to the other models, as it falls to rank 33 on the x-axis, while the blended, WMA, and SMA models rank in positions 6, 9, and 13 in validation, respectively. In both calibration and validation, the blended model ranks higher by average rank than the HMETS or averaged models.

The results indicate a successful application of the simultaneous calibration of parameters and model structure, consistent with some of the best available fixed model structures and averaged models in this study in both calibration and validation performance. This indicates that the blended model is successfully deployed with some evidence that the blended model has the potential of enhancing the overall performance of hydrologic models. In the comparison of the blended model calibration NSE rank within each catchment (i.e. **Figure 3.2**), the blended model is shown to have a robust performance across the twelve hydrologically varying MOPEX catchments, including the top performance in a number of catchments. This high performance of the blended model is seen even with a larger number of parameters than all fixed models. The robustness is emphasized in the ranked plot in **Figure 3.3**, where the blended model ranks as the highest in calibration for both calibration schemes, and

outperforms many of the fixed structures in validation as well. The relative performance of HMETS versus that of the blended model is revealing. HMETS consistently performs very well in all calibrations. This, in part, reflects the quality of the named model structure, but also may indicate that the alternate model structures (i.e., the algorithms not found in HMETS) either have reduced flexibility or appropriateness for these catchments. Likewise, the alternate model structure set may have been insufficiently diverse or independent, thus enabling the calibration of HMETS to focus on fine tuning a much smaller set of fixed parameters while the calibration of blended model incurs significant overhead in trying to simultaneously identify model structure and tune parameters. The dependence of the blended model performance on the chosen variety of structures or different calibration budgets is not evaluated here.

In this study the 108 fixed model structures represent all possible non-blended permutations of the available model structures. A requirement of the blended model to outperform all 108 fixed structures across all catchments and independent trials, may be a superficially high standard to demonstrate an improved performance over fixed model structures. However, the blended model structure is shown to generally within the top quartile of model structures when examining the maximum NSE within each catchment, and furthermore is demonstrated to be the top ranked model structure on average among 108 fixed structures, where the 108 fixed structures represent the discrete boundaries in model space to the blended model.

The comparison of the blended model structure to the averaged model structures is

also a rigorous one, where the averaged models represent ensembles of 108 fixed model structures calibrated separately and cumulatively involves a much higher computational cost to generate. The use of model ensembles is currently a common method in hydrology to combine model outputs and sample model space (*Remmers et al.*, 2020). In this study the blended model is found to outperform the averaged models in almost all instances, while requiring less than 1% of the computational cost. For example, our benchmarking shows in a single blended model simulation over the calibration period, the blended model run time only increases by just over 2% relative to the average model executable run time for the 108 fixed structure models in our experiment. This minor run time increase is a major advantage of blending at the process level instead of at the model output level. Given the negligible runtime increase of the blended model versus other models, in this setup with a fixed calibration budget for each model, the averaged models require 10 000 model runs for calibration of each of the 108 fixed model structures, while the blended model requires only one calibration with 10 000 model runs (i.e. approximately 1/108 the computational cost). Even if the calibration budget of the blended model would need to be doubled in order to account for the additional parameters and longer convergence time, this would be a significant reduction in the computational budget of the experiment overall. Rarely, if ever, would a given model structure be anticipated to outperform more than one-hundred related and individually calibrated model structures, or the ensemble of those individually calibrated models.

Additional insight on the increasing validation NSE scores with lower calibration

NSE scores in particular catchments, such as in catchment 11 (**Figure 3.2**), may be gained from the plots of observed hydrographs for each catchment (**Figure 7.1**). In catchment 11 there exists a disproportionately large peak flow event that occurs in the calibration period, and no event of a similar magnitude can be found in in validation. In this case, the calibrated models may be overfitted to this event, particularly since the usage of NSE will tend to emphasize fitting to large flow events (*Schaefli and Gupta*, 2007). This may cause the calibrated models to be overfitted to large flow events in calibration, and underperform in validation when no such large events exist in the observed data. Some of the same discrepancies in calibration and validation scores for catchments 11 and 12 in particular (Guadalupe River TX, and San Marcos River TX, respectively) are found in other studies using the MOPEX12 dataset (e.g., *Ye et al.*, 2014; *Spieler et al.*, 2020; *Clark et al.*, 2008). *Clark et al.* (2008) examined the Guadalupe and French Broad catchments, and suggested that "some structures may be better suited than others" in the arid Guadalupe catchment, while the structures had enough flexibility to be provide good performance in the wetter basin.

The results of the maximum calibration NSE for each catchment (**Figure 3.2**) demonstrate a high degree of equifinality in a number of cases, where many model structures exist with similar NSE scores. In **Figure 3.2**, many catchments (e.g. catchments 2, 3, 4, 5, 7, 9, 10) have highly similar NSE scores among the first 30 or more model structures, and any selection of a dominant structure based on the single NSE metric would be ill-advised. For example, in catchment 9, the blended and HMETS

models are in rank positions 22 and 29, respectively, but the difference in maximum calibration NSE to the top ranked model in catchment 9 is less than 0.005 NSE for both models. In this catchment, the 37 top ranked models are virtually indistinguishable from one another based on calibration NSE performance alone.

One possibility for addressing this equifinality in the calibration performance is to examine the validation performance of the models, although it is not uncommon (for example, catchments 4 and 7 in **Figure 3.2**) to find similar performance metrics for these model structures in the validation period as well. Another method for addressing equifinality is the use of multiple metrics to discriminate between model performances, such as the use of hydrologic signatures (*Euser et al.*, 2013; *Shafii et al.*, 2017; *Kavetski et al.*, 2018), and the use of multiple other metrics (e.g. *Knoben et al.* (2020)). However, the purpose of this study is not to discriminate between model structures and select the optimal structure *per se*, but to demonstrate that the calibration of a blended model structure is a viable alternative to calibrating multiple fixed model structures. Therefore, equifinality in the single NSE score is acknowledged but does not undermine the conclusions drawn from the presented results.

An application of the blended model is the possibility of identifying dominant processes and model structures, which may be done through the calibration of the blended model and without the enumeration of all plausible fixed model structures individually. As demonstrated in the results presented in this section, the blended model structure may select blends of model structures (in effect, seeking new model structures) in

cases where a single dominant structure or process does not exist within the provided framework. The identification of dominant process algorithms and model structures is discussed further in **Section 3.3.2**.

3.3.2 Model Selection Performance

In this section, the ability of the blended model to 'select' preferred model structures at the individual process level (i.e., for baseflow, evapotranspiration, infiltration, quickflow, and snow balance) is evaluated with reference to the 'brute force' approach of calibrating 108 fixed model structures. Model structure selection, in the case of the set of 108 fixed structure models, corresponds to individual structures consistently performing better than their alternatives across the range of models. In the case of the blended model, model structure selection is indicated by a preference to weight one structure over the alternatives. The latter approach is clearly less computationally intensive, and it is therefore worthwhile to evaluate whether it is as effective.

The model performance results based on a calibration using a 10 000 budget are provided in **Figure 3.4**, and for brevity are shown only for catchments 1, 8, and 11. The three catchments are selected as illustrative examples in the relative performance of the blended model in different catchments, determined from an examination of the process plot in **Figure 3.4**, as well as **Figure 3.2**. These three catchments were each deemed representative of the general characteristics of a subset of the 12 MOPEX catchments as follows:

- Catchment 1: the HMETS model generally has a higher NSE score than the blended model, with no clear correlation between the selection of process algorithms and overall model performance in calibration, with the exception of the selection of the infiltration algorithm (where infiltration algorithm no. 1 is clearly preferred). Catchment 1 is considered representative of catchments 3, 5, and 10 in this respect.

- Catchment 8: the HMETS and blended models perform similarly well in calibration, with a clear separation in model performance based on the selection of the snow balance algorithm, and the preferred infiltration algorithm also corresponds to process option no. 1. Catchment 8 is considered representative of catchments 4, 6, and 8 in this respect.

- Catchment 11: the blended model has a higher NSE score in calibration compared to the HMETS model, where the calibration performance is clearly reliant on the selection of the non-HMETS evapotranspiration algorithm (process option no. 2) and the selection of infiltration algorithm no. 1. Catchment 11 has been described in other studies as an arid catchment where the selection of model structure appears to be a more sensitive decision for NSE performance than than in wetter catchments (e.g. *Clark et al.* (2008)).

The results for all twelve MOPEX catchments are provided in the **Appendix**.

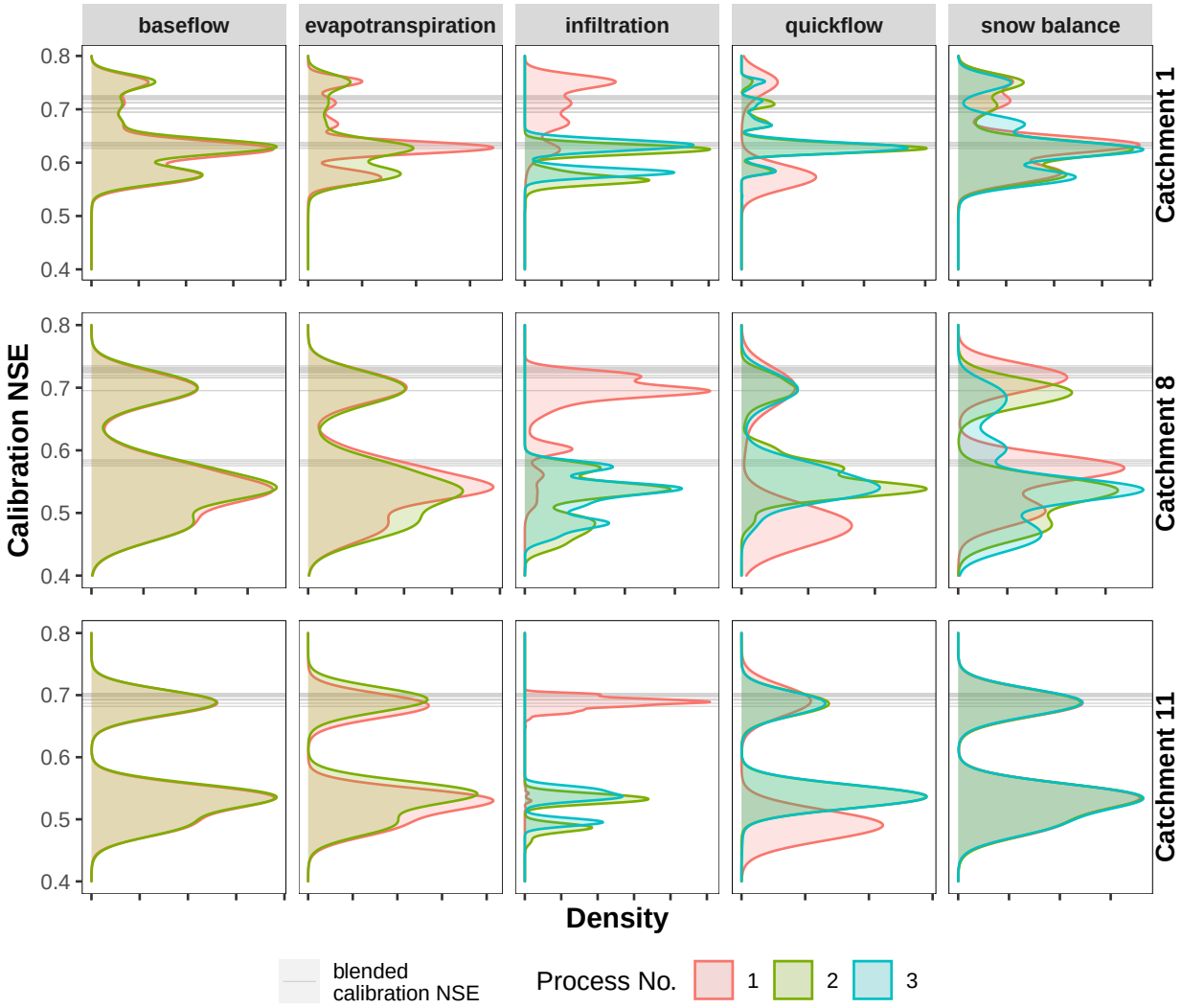


Figure 3.4 Model performance in calibration NSE score on the y-axis and density on the x-axis (1-109), plotted for catchments 1, 8, and 11, and faceted horizontally by the five groups of process algorithms in this study. Density axis ranges vary by plot but labels have been removed for simplicity. In each density curve the colour corresponds to a choice of process algorithm within the process group as 1, 2, or 3. Model structures 1 through 108 are fixed model structures. The horizontal grey lines indicate the NSE performance of the blended model trials. The averaged model is omitted from this analysis. Results are shown for a calibration budget of 10 000 model runs.

Figure 3.4 is initially examined only with reference to the distributions generated using the brute force approach of model selection. It is observed that in all catchments that the HMETS infiltration algorithm (process option no. 1) is required for a high NSE score in calibration. This is indicated by a distinct distribution of higher NSE when process no. 1 is used. This may be contrasted with the baseflow algorithm, where the distributions of NSE for all three catchments are nearly indistinguishable for the two process options. This infiltration algorithm is clearly preferred for the three catchments in **Figure 3.4** and the remaining nine MOPEX catchments (**Figure 7.2**).

In catchment 1, the separation between the three snow balance algorithms is not clear, with a similar performance seemingly achieved with any of the three snow balance algorithms. In catchment 8 there seems to be a stronger preference for the HMETS snow balance algorithm (process option no. 1), with a secondary preference of the third snow balance algorithm over the second. This preference is clearer when examining only NSE above about 0.65 (i.e., when conditional upon the proper selection of infiltration algorithm). In catchment 11, which is a Texas catchment with no snowfall in a typical year, there is a clearly no effect of the snow balance algorithm on model performance, as would be anticipated.

Catchment 11 shows a slight preference in the second evapotranspiration algorithm over the HMETS evapotranspiration algorithm, unlike catchments 1 and 8 with no clear difference in model performance based on the evapotranspiration algorithm. The higher importance of the selection of the evapotranspiration algorithm on model per-

formance in arid catchments than others is also anticipated. This preference for the non-HMETS evapotranspiration routine in catchment 11 may be part of the reason why the HMETS model does not perform as well in this catchment, and performs worse in the calibration performance ranking in this catchment than in any other. The full process results (**Figure 7.3**) however, show that this preference for the non-HMETS evapotranspiration algorithm is less clear in catchment 12, the second Texas catchment. Therefore the dominance may be for a more local reason than as an arid Texas catchment behaviour overall.

A preference for the HMETS quickflow algorithm in catchments 8 and 11 may exist, although it is clearly shown for catchment 1 where the HMETS algorithm (option 1) dominates the upper tail of the performance density curve. There appears to be little impact on the model calibration NSE score based on the selection of the baseflow algorithm in these three catchments; this is likely a function of the objective function, which is less influenced by low flow characteristics of the hydrograph.

The process weights from the blended models are calculated using the calibrated weight-generating parameters for each independent trial. The process weights for each blended process in catchments 1, 8, and 11 are plotted for the twenty independent trials in **Figure 3.5**. These results are filtered to use only independent trials where the calibration NSE score is within 0.05 of the top performing model, which removes trials with poor scores that may not be reliable for structure identification (i.e, under-optimized solutions). The full set of plotted process weights for all twelve catchments

is provided in the **Appendix** (**Figure 7.3**). Here, it is implicit that higher weights associated with a process correspond to model structure selection. Blended model selection skill is evaluated with respect to the brute force results of **Figure 3.4**: if the same preferred model structures are identified, then the blended model selection is deemed to be effective.

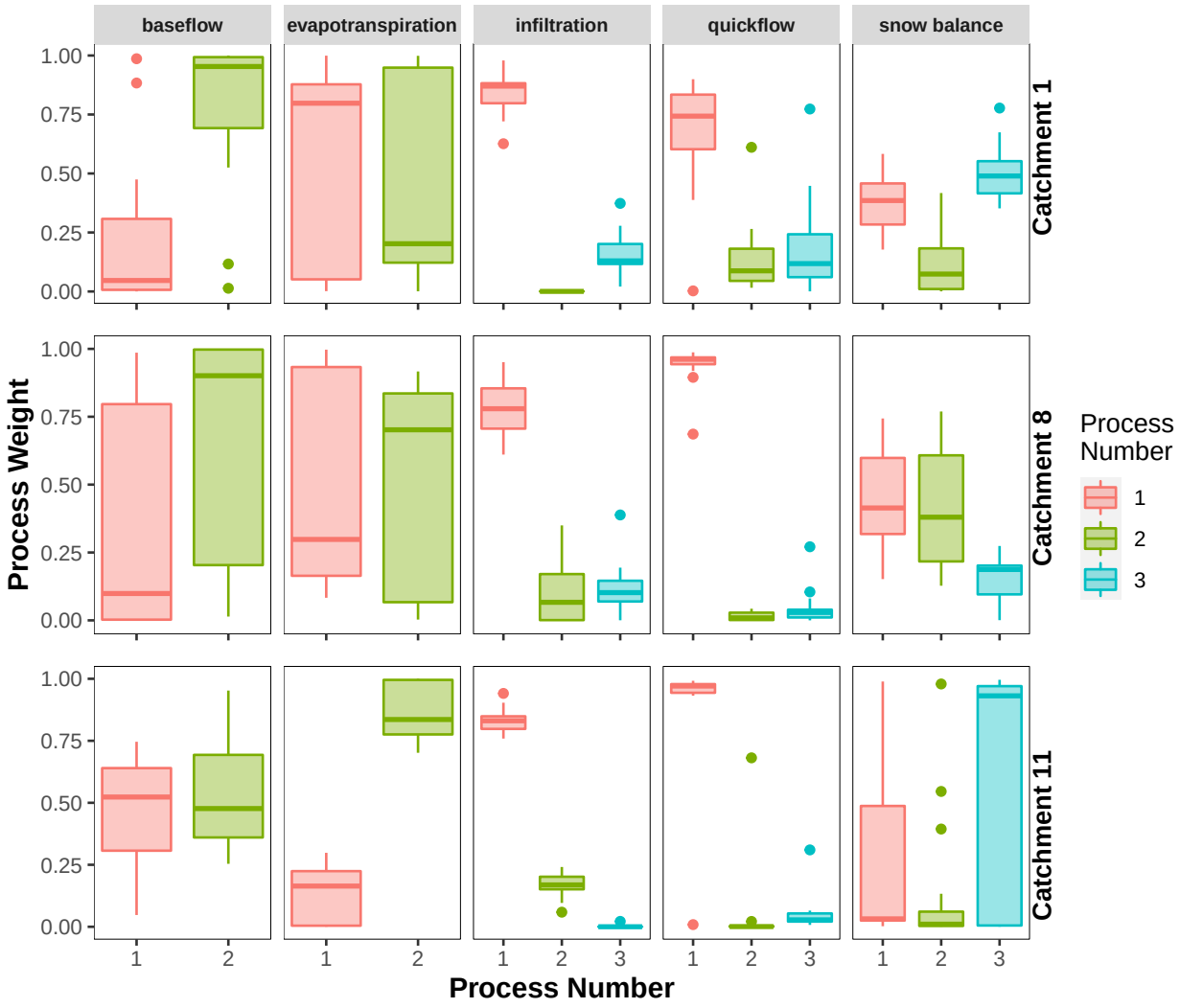


Figure 3.5 Blended model process weights from 0 to 1 on the y-axis, and numbered options for each blended hydrologic process on the x-axis, faceted by the hydrologic process group. Each box and whisker plot displays the spread of the given process weight within the process group, where horizontal lines represent the 25^{th}, 50^{th}, and 75^{th} percentiles, and outliers are shown as points. For each catchment, the data blended model result is filtered to include only information from independent trials where the calibration NSE performance was within 0.05 inclusive of the top performing independent trial. Results are shown for a calibration budget of 10 000 model runs, and only catchments 1, 8, and 11 of the MOPEX12 catchments.

In catchments 8 and 11 the baseflow algorithm in the blended model has approximately equal weighting on average, with no clear pattern in performance related to the selection of one baseflow algorithm or the other, which is consistent with the fixed model results (**Figure 3.4**). All of the clearly preferred options identified in the fixed structure evaluation are likewise reflected in the blended model weights. The first infiltration algorithm is strongly favoured in all three catchments. The evapotranspiration algorithm shows approximately equal weighting on average in catchments 5 and 8. However, it shows a higher process weighting for the second algorithm in catchment 11, as is consistent with the slight preference shown in the fixed model structures. The quickflow process weights tend to favour the first quickflow algorithm in catchments 8 and particularly 11, with a less dominant but still clear preference in catchment 1. Finally, the snow balance is shown to have a slight preference a blend of algorithms 3 and 1 in catchment 1, with a low blended weight assigned to algorithm 2.

The blended model calibration did seem to select some algorithms not clearly shown to be preferred in the evaluation of the 108 model structures. For instance, in catchment 1 there is a preference for baseflow algorithm 2 with the blended model, which is not seen in the fixed model structure results of **Figure 3.4**. This may be because of the relatively small impact of low flows on NSE or could be connected to the relative low ranking of the blended model in this catchment, perhaps implying this model suffers from this preference during the calibration process. In catchment 11 there appears to be an inability to differentiate between snow balance algorithms 1 and 3, with a

handful of outliers that more strongly favour algorithm 2. With the understanding that the snow balance algorithm selection should not impact the model performance in a Texas catchment, this may indicate that the algorithm may weight processes arbitrarily, and therefore a high weighting alone may not be an indication of preference; rather, consistently high weighting may be more informative. In relation to the fixed structure results, the snow balance algorithm preferences in catchment 1 are somewhat different from the fixed model structures, where all three algorithms appear to have similar performances, and the blended model weights skew towards algorithms 1 and 3 more heavily. In catchment 8 the preference for algorithms 1 and 2 aligns between the fixed structure results and the blended model weights more clearly, and the lack of preference in catchment 11 is also consistent between the two sets of results. Additional discrepancies may be seen with respect to quickflow processes in catchments 8 and 11, in which the blended weights demonstrated no clear preference but there is a difference in the NSE distributions of the 108 models. Of course, neither the weights or NSE performance distributions are unassailable indicators of model structure preference; their quality is conditional upon the number of independent trials used, the convergence of the calibration, and the equifinality of model structure. Because of this, we are wary of expecting uniform agreement between the two metrics, which are both potentially useful indicators of submodel preference that deserve future investigation.

In examining the blended model weights for all twelve catchments (see the **Appendix**), the following results are observed:

- the baseflow process algorithms tend to have near-equal weighting in almost all catchments;
- the evapotranspiration process algorithms tend to have similar weights in most catchments, with some exceptions where one of the algorithms is more heavily weighted;
- the first infiltration algorithm (HMETS infiltration) appears to be preferred in most catchments, with exceptions in catchments 2 and 12, where it is equally weighted with one of the other two algorithms;
- the quickflow algorithm tends to skew towards the first process algorithm, but is more equally weighted with other options in many catchments; and
- the snow balance algorithms have a variety of preferences across the catchments for one of the three algorithms, a blend of two algorithms, as well as no clear preference between the three algorithms.

In cases where no clear or less clear dominant process algorithms tend to exist, as shown in some cases with the with baseflow and evapotranspiration algorithms, it is unclear if the inclusion of additional process algorithms will eventually reveal a dominant process algorithm within the group, or the addition of more process algorithms to the group will fail to identify a dominant one and the choice of algorithm for the given process is simply less important than it is for other process groups. For example, the process groupings for baseflow and evapotranspiration with only two process algo-

rithms appear to have less preference in the blended weights for a particular algorithm than other processes with three algorithms, possibly suggesting that two is at least one option too few. As previously noted, the choice of calibration metric likely has some influence, as the lack of baseflow algorithm preferences may be due to the use of a metric biased towards high flow events. Another possibility is that the blended model may identify preferred solutions in continuous model space where the discrete model structure options do not have a clear dominant structure, which is a likely advantage of the continuous exploration of the BMSC method over existing ones. In any case, there exists a balance between including additional process algorithms and limiting the number of blended model parameters in the calibration. The appropriate number of process algorithms to include in a grouping is not clear; more work is required to determine the proper strategy for seeding the BMSC with the most appropriate structure population.

A larger scale exploration of this type of process plot for a larger catchment set, such as the CAMELS dataset (*Addor et al.*, 2017) with its 671 catchments, may help to generalize these insights for particular sets of process algorithm selections. The potential role of the blended model in this type of analysis may be in 1) exploring various model structures with a methodology that replaces the enumeration of multiple structures, as the blended structure may emulate these structures without the requirement for separate model configurations, and 2) in exploring inter-structure model performances and interactions between process selections, i.e. exploring model space continuously.

A clustering of model performance and process weights may provide additional insights on the dominance of particular processes and process algorithms. This serves as a computationally-efficient potential alternative to the quantification of fixed model structures for classification of model structure families, such as done by *Remmers et al.* (2020).

3.3.3 Impact on Model Calibration with Non-Random Seeding

In the results presented thus far, the blended model is calibrated using a random initial parameter values in the 20 independent trials for each catchment. The third research question of this work is to explore the impact of non-random seeding on the blended model. This is discussed briefly in this section.

As an additional experiment, the blended model is calibrated using the calibrated HMETS parameters within the given trial for each catchment. The blended model structural parameters are set to emulate the calibrated HMETS model, and all non-HMETS parameters are set to default values in the initial seeding of the blended model. The blended model is calibrated with the same calibration budget (10 000 model runs) using the non-random starting values. In this sense, it may be considered to have twice the computational budget used relative to the other model structures. This blended model configuration is referred to herein as the blended$_{\text{HMETS}}$ model. The HMETS model is selected as an initial value set for this experiment due to its high NSE scores in multiple catchments, as well as its existence as the basis for many of the derived process

options in the modelling configuration. It is noted that in this experimental setup, the HMETS model structure is a subset of the blended model structure, which allows the emulation of the calibrated HMETS model with the blended model structure. The main objectives of this are to examine how this seeding would impact the performance of the blended model relative to random seeding of initial parameters, and how much improvement may be observed in a calibrated fixed model structure when the structure is switched to a blended model.

For brevity, the three catchments discussed in the **Section 3.3.2** are selected once more to examine these results. In calibration, the performance of the blended, blended$_{\text{HMETS}}$, and HMETS models are presented in **Figure 3.6**.

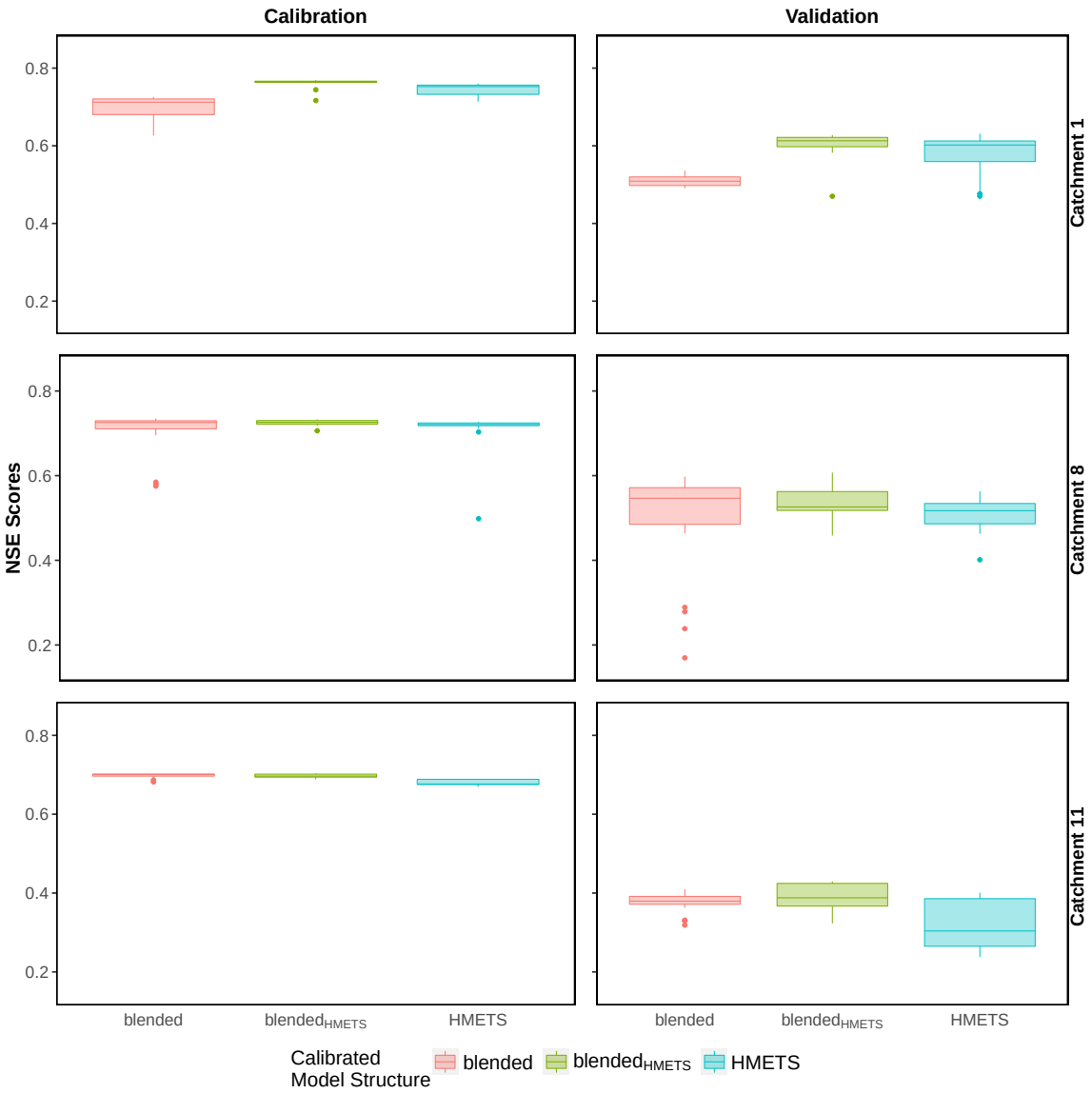


Figure 3.6 Calibration and validation NSE scores plotted for catchments 1, 8, and 11, for the blended, blended$_{\text{HMETS}}$, and HMETS models. Column 1 presents the calibration NSE scores, and column 2 presents the validation NSE scores. Catchments results are presented by row, with catchment 1 information in row 1, catchment 8 in row 2, and catchment 11 in row 3. The blended model is shown in red, the blended$_{\text{HMETS}}$ is shown in green, and the HMETS model is shown in blue.

In catchment 1, there is an improvement in the calibration NSE scores in the blended$_{\text{HMETS}}$ model over the HMETS model, as well as higher scores overall in the blended$_{\text{HMETS}}$ model over the blended model. This improvement can likely be attributed to the initial high performance of the HMETS model used in seeding. In catchment 8, the blended model with random seeding appears to generally outperform HMETS and perform similarly to the blended$_{\text{HMETS}}$ model. In catchment 11, the calibration NSE score of the blended model is generally slightly better than that of the HMETS model, which is not the case in catchments 1 and 8. In validation, the blended$_{\text{HMETS}}$ model tends to have higher NSE scores than the blended model, and both have higher validation NSE scores than the HMETS model.

The average improvement in the NSE score for the blended$_{\text{HMETS}}$ model relative to the blended model across all catchments and iterations is found to be 0.034. This indicates a marginal improvement in the calibrated NSE scores of the blended model structure when seeded with the HMETS model, on average, but likely not one that is sufficiently high to consider the blended$_{\text{HMETS}}$ model a better model overall. Based on the calibration scores alone, the blended$_{\text{HMETS}}$ model improvement also appears conditional on the initial NSE scoring and suitability of the HMETS model to the given catchment. From **Section 3.3.2**, the experiment provides some evidence that the HMETS evapotranspiration algorithm may not be ideal for catchment 11. In this case, the blended model tends to have a higher calibration NSE score overall than the seeded blended$_{\text{HMETS}}$ model, suggesting that seeding the blended model with a

calibrated initial solution may limit the ability to explore other model structures with the DDS algorithm. In the other catchments where the calibrated HMETS model has a higher calibration NSE score than the blended model initially, further calibration of the model as a blended structure continues to yield a higher scoring model than the randomly seeded blended model.

The validation NSE scores indicate a similar relative difference in model performance between catchments 1, 8, and 11, where the blended model has a lower validation NSE score than the blended$_{\text{HMETS}}$ model in catchment 1, has a slightly better validation score than the blended$_{\text{HMETS}}$ model in catchment 11, and has more mixed results in catchments 8. The overall worse NSE scores in validation than calibration are also noted, particularly in catchment 11.

The results overall indicate that the seeding of the blended model with a well-performing model, such as HMETS, conditionally improves the final NSE scores of the blended model relative to the use of random initial values. In cases where the HMETS model has a high calibration NSE score in the given catchment and the HMETS algorithms are well-suited to the representation of hydrologic process (such as catchment 1), the initial seeding leads to a marked improvement over the blended model NSE scores, and a smaller but still considerable improvement over the HMETS model NSE scores. In catchments where one or more HMETS process algorithms may not be well-suited to the catchment, such as in catchment 11, the blended model with random values scores similarly well or higher than the blended$_{\text{HMETS}}$ model, even though the

blended$_{\text{HMETS}}$ model can be considered to have twice the computational budget used in its calibration (once for the calibration of the HMETS model, and again for the blended$_{\text{HMETS}}$ model). It is noted that this appears to occur even with the use of the DDS algorithm, which is designed to search for global solutions efficiently (*Tolson and Shoemaker*, 2007).

3.3.4 General Discussion

The BMSC method shows the potential for informing the selection of preferred process algorithms and model structures through the explicit calibration of continuous model structure during the calibration process. This method is clearly less computationally expensive than a comparable brute force approach which enumerates all plausible model structures for evaluation and selection. In practice, the *a priori* knowledge of an experienced hydrologist may inform the selection of the model structure and reduce the burden in enumerating hundreds of model structures. However, a key benefit of the BMSC approach is that the blended model may identify and inform the selection of this 'optimal' model structure with relatively minimal effort. The experience of the hydrologist can be focused on deciding which groups of process algorithms would be best to include in the blended model structure, without the need to explicitly determine *a priori* which combination of these algorithms is the optimal one. The typical model-building and calibration tasks requiring the hydrologist's attention, including model discretization, providing data inputs, determining the calibration budget and

performance metrics, are still included in the blended model-building process.

The current BMSC method addresses some of the efficiency issues in the mixed-integer AMSI approach introduced by *Spieler et al.* (2020) in which inactive parameters are still calibrated, by making all parameters active in the calibration. However, an inefficiency in this approach still exists in terms of the calibration budget expended by calibrating the less active parameters; for example, when the blended model weighs a particular process to nearly zero, the parameters tied to that specific algorithm with near zero influence on the model performance are still considered 'active' in the calibration. The results herein indicate that the current configuration is capable of achieving a 'good' performance, but with convergence on a solution possibly requiring more than 10 000 iterations to fully converge in some catchments using DDS. This suggests that a more efficient calibration method may be desirable for calibrating the blended model.

One limitation of this study is the use of the single integrated NSE metric. The NSE provides some indication of the model performance but marginal differences in the NSE metric are insufficient to inform the overall relative performances of the models (*Gupta et al.*, 2008). The consistent use of the NSE metric for all models in this study is considered sufficient for the purpose of this study, which is to present the BMSC method as providing, at a minimum, comparable model performances to related fixed model structures. However, additional metrics in the calibration and description of the models must be used to discriminate between models with a high degree of equifinality

if a 'best' model is to be selected, and multiple metrics (such as hydrologic signatures) are recommended for an operational or scientific deployment of the BMSC method. The technique used here, however, is readily extended to other forms of objective function.

Another limitation in this study, and direction for future studies, lies in the determination of the appropriate number and types of process algorithms to be included in a given group of process algorithms (as discussed in Section 3.3.2). In this study we use two or three algorithmic options within several process groups. However, additional effort in determining the correct number of process algorithms with sufficiently varying performances to maximize the robustness of the blended model structure to various catchments would improve the BMSC performance in future applications. We expect that the BMSC will be most robust when seeded with process algorithms that are sufficiently independent and cover a large range of model structure space. Reconciling the blending of highly conceptual process algorithms with more physically-based algorithms may also provide a direction for future work.

While the blended model structure is applied to lumped representations of catchments in this study, the same set of process groupings can be directly applied in a distributed context, where the parameters associated with each algorithm may vary by soil class, land use class, etc. The flux for each grouped process would be calculated in the same manner but done so within each HRU, where the flux within the HRU is now a function of the parameters associated with the properties of the given HRU. The individual fluxes for a given algorithm will therefore vary within each HRU based on

the algorithmic parameters provided, but the same weight-generating parameters may be used to blend these fluxes, and no additional weight-generating parameters would need to be introduced to the model. A similar set of analyses to assess model performance and identify dominant model structures could be done in a distributed context, since the parameter values of non- weight-generating parameters were not considered. Additional data to justify the use of distributed modelling may also help to constrain the optimization of weight-generating parameters (e.g., snow depth measurements at key locations may improve the identifiability of snow balance process algorithms).

3.4 Conclusions

The simultaneous calibration of hydrologic model structures and parameters is a new strategy in hydrologic modelling (*Spieler et al.*, 2020), and is advanced in this study with the introduction of the Blended Model Structure Calibration (BMSC) method, which continuously samples model space between discrete model structures. The BMSC method is shown to provide robust calibration NSE performances across twelve of the MOPEX catchments. The blended model structure calibration results are overall better in comparison to 108 fixed and individually calibrated model structures and their averaged output. The fixed structures represent all possible permutations of model structures within the set of process algorithms included in this study. The blended model structure is ranked as the top performing model structure based on the NSE performance after calibration when examining average ranks across the twelve catch-

ments and twenty independent trials performed. The ability of the BMSC to identify preferred process algorithms was also shown through the selection of weight-generating parameters in the calibration process, and generally identifies similar preferred model structures as are found via evaluation of all fixed model structures, while using a significant fraction of the computational cost.

The BMSC method continues along this paradigm shift in the model-building process, as the feedback of information from the calibration process informs the selection of the model structure. This is a distinct change from building a conceptual model and selecting the process algorithms *a priori*. The BMSC formulation builds on the work by *Spieler et al.* (2020) by enabling modellers to calibrate model structure using the same continuous-variable optimization algorithms that have historically been applied for hydrologic model calibration. It is anticipated that the BMSC method will prove useful for undertaking scientific studies to understand the role of model structure in describing watershed hydrological function, and for improving the performance of hydrologic models for operational usage.

Chapter 4

Optimizing a general hydrologic model structure

Statement of Contribution

The following chapter is written based on the following article (*Chlumsky et al.*, 2024a):

Chlumsky, R., Mai, J., Craig, J. R., and Tolson, B. A. (2024): Advancement of a Blended Hydrologic Model for Robust Model Performance. Journal of Hydrologic Engineering, 29 (5), 04024,033. doi:https://ascelibrary.org/doi/10.1061/JHYEFF.HEENG-6246.

This chapter presents a modelling study in which the blended model structure is experimented with by altering the algorithms and process options, including new blended forcing groups. This chapter presents the development of version 2 of the

blended model structure. In this chapter, Robert Chlumsky designed the modelling experiment, wrote the code to efficiently carry out the experiment, analyzed the results and prepared the chapter including handling of revisions during the review process. Drs. Mai, Craig, and Tolson previously published the blended model structure deployed in this study (*Mai et al.*, 2020), were co-authors on the preceding work by *Chlumsky et al.* (2021), and contributed to the interpretation of results and revisions of this chapter.

4.1 Introduction

Hydrologic models have been useful tools in the hands of hydrologists for many decades. The exponential increase in computational power available to researchers and practitioners is partially responsible for the rapid development of hydrologic models, where discussion around the 'plethora of hydrologic models' (*Clark et al.*, 2011) and the existence of too many hydrologic models (*Horton et al.*, 2021) have been noted in more recent literature. *Weiler and Beven* (2015) advocate for the development of a so-called Community Hydrologic Model, which would have enough flexibility to offset the increasing number of new hydrologic models being developed. Such a model could be applied at different scales and for different watersheds without a need for developing new models. The hydrologic modelling community is still largely in search of a model that successfully fulfills this purpose, though for many researchers the focus is now on understanding when and where to deploy particular models instead of creating a single Community model.

Despite this, the number of models being developed and used in studies is increasing. *Nearing et al.* (2021) state that the reason for the relatively unsuccessful efforts to nonetheless find scale-relevant theories in hydrologic models is simply that the hydrology community has failed to find them. *Nearing et al.* (2021) argue that the ability of deep-learning models to capture complex relationships in catchments, and to provide better daily streamflow predictions in ungauged basins than traditional hydrologic

models do in gauged basins demonstrates that these relationships can be found, but have not been discovered through traditional modelling efforts. The traditional belief that overparameterizing a model leads to poor validation performance also does not appear to be the case with deep-learning models (e.g., see *Mai et al.*, 2022a)), which tend to have orders of magnitude more parameters than process-based hydrologic models. Additional work in the machine-learning community is being done to improve the realism of machine-learning and deep-learning based models, such as the inclusion of mass-balance constraints on models (*Frame et al.*, 2023), addressing a common criticism of machine-learning approaches. The relatively successful efforts of deep-learning models in hydrology (*Nearing et al.*, 2021; *Nevo et al.*, 2022; *Frame et al.*, 2022; *Lees et al.*, 2022; *Klotz et al.*, 2022; *Arsenault et al.*, 2023) leave the process-based modelling community with a fundamental question of how process-based models can be made to capture complex streamflow signals, as demonstrated with deep-learning models. The recent success of machine-learning and deep-learning models relative to process-based or conceptual models in controlled intercomparison studies (e.g., *Mai et al.*, 2022a)) further encourages hydrologists pushing for process-based approaches to confront the degraded predictive accuracy of their models. Recent studies have sought to understand the success of deep-learning models, and the connection to process-based models (e.g., *Kratzert et al.*, 2019b; *Frame et al.*, 2023)).

Here, we argue that the recently introduced blended model approach (*Mai et al.*, 2020) may serve to provide some answers to these important questions in process-based

hydrologic modelling, as both a plausible basis for the Community Hydrologic Model and as a substantial step forward in process-based hydrologic model development. A blended model is one that uses multiple process options to determine the flux of a hydrologic process (e.g., infiltration) within each time step of the model; model weights can be adjusted through calibration or other exercise, which effectively provides the flexibility in the model to change its structure without a need for separate model codes or setups. This is a feature that may allow a well-constructed blended model configuration to serve as a form of generalized model applicable to different scales and catchments. This characteristic of blended models also allows them to not only maintain a relatively high performance across a range of catchments, but inform the selection of preferred structural options, and do so with a fraction of the computational cost relative to the more common approach of running multi-model ensembles, as demonstrated in *Chlumsky et al.* (2021). The additional complexity of process representation introduced by the blended modelling approach, while maintaining the important features of process-based models such as mass conservation and physical interpretation of state variables, is a feature that we suggest may allow for the successful representation of complex watershed dynamics that are seemingly captured by deep-learning models. In the recent Great Lake Runoff Intercomparison Project (GRIP) over the Great Lakes watershed (i.e., GRIP-GL), the original blended model of *Mai et al.* (2020) was one of the most successful among process-based models in a comparison against the dominant deep-learning model (*Mai et al.*, 2022a).

In this study, we iterate upon the original blended model introduced by *Mai et al.* (2020) and employed by *Chlumsky et al.* (2021), *Mai et al.* (2022b), and *Mai et al.* (2022a) with the goal of producing a generally improved blended model. This goal is distinct from building an optimal model for a single application and catchment; instead we aim to design a blended structure that can be robust across many applications and catchments. Since the original blended model configuration has been adjusted since these initial deployments, we anticipate that iterative experiments to its structure and process equations may yield some overall improvements. The process undertaken in developing a new configuration for the blended model is also discussed in this work, and may be useful in future development of blended models.

The main objectives of this study are to (1) test additional blended model configurations, (2) improve upon the original blended model by creating a new blended model version that is tested in multiple catchments, and (3) validate the updated blended model version with additional catchments not used in objective (2) to demonstrate its improvement.

The remaining sections are organized as follows. The methods (Section 4.2) is organized into several sections that discuss the blended model background, the iteration on model configurations, and the eventual selection and validation of a preferred model configuration. Results and Discussion (Section 4.3) are presented in parallel, presenting the results associated with each method described along with accompanying discussion of those results. Concluding points are presented in Section 4.4.

4.2 Methods

The following methodology sections are organized as follows. **Section 4.2.1** introduces the blended modelling approach and provides the necessary background on the original blended model. **Section 4.2.2** discusses the iterative development and evaluation of candidate blended model configurations. **Section 4.2.3** examines the methodology undertaken to assess model overfitting, and ensure that additional model parameters do not decrease the performance of blended model configurations in validation. **Section 4.2.4** discusses the selection process for a new version of the blended model from the set of plausible blended model configurations. Finally, **Section 4.2.5** describes the process to validate the new blended model version with a set of catchments independent from the development and selection of the new blended model version.

4.2.1 Blended hydrologic model

The concept of a blended model was introduced by *Mai et al.* (2020), and provides a method to include multiple process options for use in calculating hydrologic process fluxes within a single model simulation. A given blended model configuration defines all model state variables, processes and process options that can be simulated. The original blended model configuration included five process 'groups': infiltration, quick-flow, baseflow, evapotranspiration, and snow balance. Within each process group $\mathcal{G}$, the process flux (e.g., infiltration rate) for that process at a given timestep t, $f_{\mathcal{G}}(t)$, is

calculated as a weighted average of the model output of $NP_{\mathcal{G}}$ process options. Mathematically, this may be expressed for any given process group as:

$$f_{\mathcal{G}}(t) = \sum_{i=1}^{NP_{\mathcal{G}}} w_{\mathcal{G}i} f_{\mathcal{G}i}(t) \tag{4.1}$$

where $f_{\mathcal{G}i}(t)$ is the process flux f (typically in mm/d) simulated for time step t by the i^{th} process algorithm within the group $\mathcal{G}$, and $w_{\mathcal{G}i}$ is the weight associated with the i^{th} process algorithm within the group $\mathcal{G}$ (here, weights are time-invariant). Parameters associated with two or more process algorithms may be shared in the blended model. For example, if process algorithm 1 within a given group $\mathcal{G}$ is of the form $f_{\mathcal{G}1} = k_1 * S$ and process algorithm 2 is of the form $f_{\mathcal{G}1} = k_2 * S^n$, where k_1, k_2, and n are parameters and S is a state variable, then the blended model may be configured use a single parameter k_1 (i.e., $k_1 = k_2$) in order to reduce the number of parameters in the model.

This blending approach is implemented in the Raven hydrologic modelling framework (*Craig et al.*, 2020), and can be extended to hydrologic process where multiple process algorithms are available. Raven is an open-source, object-oriented software framework with more than 100 process algorithms encoded, which allows for a large selection in building both blended model configurations and flexible model structures more generally. Two key design principles for the Raven software include efficient runtime and model flexibility, making it an ideal choice for this type of research where the model must be continually modified and run millions of times. The Raven Hydrologic Modelling Framework v3.8 is available publicly at `http://raven.uwaterloo.ca/Downloads.html`

In Raven, the weights for each process group may be supplied in one of two ways; either directly as weights that sum to unity, or they may be supplied as $NP_{\mathcal{G}} - 1$ independent numbers each distributed uniformly between 0 and 1, i.e., so-called weight-generating parameters. If supplied in the latter way, these weight-generating parameters are transformed within Raven using the so-called "pie share" method (*Mai et al.*, 2022c) of generating random numbers summing up to unity while making sure the random variables are independent and identically distributed. This approach of determining weights also allows the weights to be sampled independently, since the constraint of summing to unity is met as part of the transformation of these $NP_{\mathcal{G}} - 1$ weight-generating parameters to $NP_{\mathcal{G}}$ weights.

The blended approach respects the water balance, since all weighted fluxes are limited by water availability within the relevant storage units, which is the same handling as non-blended fluxes. A blended model is also consistent with the structure of other process-based and conceptual hydrologic models with respect to tracking of state variables, such as soil moisture. This approach is therefore distinguished from machine-learning approaches in which these characteristics typically do not hold. In cases where a blended process group is used to estimate the flux of a hydrologic process in a model, this should simply be viewed as another type of algorithm used to estimate that flux (albeit a more complicated one), but otherwise equivalent to swapping the equation of choice for a given process.

Theoretically, blended models allow for the exploration of model space that is not

available when only discrete process options are used in flexible modelling frameworks. Here, we define model space as the set of possible solutions for a given hydrologic model configuration(s) to achieve. The blended model allows a single model configuration to encapsulate the model space that was previously only possible with an ensemble of varying model structures. This model space is expanded further by the continuous weighting of process options that allows model structure to be expressed as a continuum of plausible options, rather than discrete points represented by separate model structures (e.g., *Spieler et al.*, 2020)). If we imagine that optimal model solutions may exist within this continuum, rather than precisely at a specific model structure defined by discrete options (as would be the case in an ensemble), then it becomes easy to imagine why the blended model structure generally performs better than individual discrete model structures, as was found in *Chlumsky et al.* (2021) and *Mai et al.* (2022a). While the exploration of model space is still limited by the practical limits of numerical algorithms, as well as the selected process options in a given hydrologic model, the blended approach allows for an efficient exploration of a much larger model space than traditional hydrologic models.

4.2.2 Blended model development and performance evaluation

This study deploys a blended model with various selections of process algorithms and model structure changes in over 30 model configurations, with up to seven blended process groups used in any given model configuration. Here, we define a model configu-

ration as a specific blended model implementation with a unique selection of hydrologic process algorithms, including those within each blended process group. A new model configuration may be made by changing the set of algorithms within a blended process group, adding or removing a process group entirely, or by adjusting the non-blended process algorithms. In the model development phase, blended model configurations are tested in calibration mode using the 12 catchments from the second and third workshop of the Model Parameter Estimation Experiment (MOPEX) (*Duan et al.*, 2006). These 12 catchments cover the southeastern portion of the United States, and were selected because they represent a relatively diverse range of hydrologic conditions, making them suitable for testing a given blended model configuration for robustness across various conditions. More information on these catchments is provided in Section 4.2.5.

In each calibration, a blended model configuration is run with a daily timestep and calibrated using the Dynamically Dimensioned Search (DDS) algorithm (*Tolson and Shoemaker*, 2007), as implemented in the Ostrich calibration toolbox (*Matott*, 2017). DDS is applied with a calibration budget of 10 000 iterations with a two year warm up period followed by a calibration period of 12 years from 1972 to 1983. In each calibration, the traditional model parameters associated with each process algorithm as well as the weight-generating parameters are calibrated together. The Kling-Gupta Efficiency (KGE) metric (*Gupta et al.*, 2009) is used as the objective function, which is calculated as:

$$
\begin{aligned}
KGE &= 1 - \sqrt{(r-1)^2 + (\alpha-1)^2 + (\beta-1)^2} \\
r &= \frac{cov_{so}}{\sigma_s \sigma_o} \\
\alpha &= \sigma_s / \sigma_o \\
\beta &= \mu_s / \mu_o
\end{aligned}
\tag{4.2}
$$

where r is the linear correlation coefficient between x_s and x_o, α is a measure of relative variability, β is a ratio of mean simulated and observed values, cov_{so} is the covariance between simulated and observed values, σ_s and σ_o are the standard deviation of simulated and observed values, respectively, and μ_s and μ_o are the mean of the simulated and observed values, respectively.

The temporal validation scores are then calculated for the six year period of 1984-1989. Each calibration was performed in all 12 MOPEX catchments independently and repeated with 20 independent calibration trials for each deployment of a blended model to account for the variable performance of the optimization algorithm. The calibration and validation setup is consistent with the approach used in *Chlumsky et al.* (2021), with the exception of using KGE here rather than NSE as the calibration objective function. The KGE metric was selected over the Nash Sutcliffe Efficiency (NSE) metric as the KGE is considered to be a better indication of overall model fit with a more balanced evaluation of high and low flows, with some studies suggesting that KGE may be a better choice even for high flows than NSE (*Mizukami et al.*, 2019). The NSE metric is defined in *Gupta et al.* (2009) as:

$$
NSE = 1 - \frac{\sum_{t=1}^{n} (x_{s,t} - x_{o,t})^2}{\sum_{t=1}^{n} (x_{o,t} - \mu_o)^2} \tag{4.3}
$$

where n is the total number of time steps, $x_{s,t}$ is the simulated model value at time t, $x_{o,t}$ is the observed value at time t, and μ_o is the mean of the observed values.

This study adopts a comparative and empirical strategy to inform model development. This approach compares all subsequent blended model configurations against previous ones, including the original blended model configuration, to check for improved performance. This type of empirical model development is different from a more traditional approach of model development via prior perceptions of dominant hydrologic processes, and has been used in the development of conceptual hydrologic models such as GR4J (*Perrin et al.*, 2003).

With the analysis of each successive blended model configuration, the primary approaches to examining the empirical results includes the use of box and whisker plots of calibration and validation performance, both in individual catchments and across all catchments. The median and maximum KGE performance is visually examined and compared between model configurations. The weight distributions of process options within each process group (such as Figure 5 in *Chlumsky et al.* (2021)), are also examined within a model configuration to determine if a given process option is a) not being selected within a given catchment, or b) not being selected in any catchments. Process options with low weights across the 20 independent trials and across all catchments (scenario b)) indicate that the process option is not a valuable contribution to the process group and therefore the given blended model configuration overall; these process options are replaced in a subsequent model configuration with others. While

a process algorithm may be removed in this experiment, this does not indicate that the algorithm itself is not useful in other model configurations, simply that it may not be useful in the particular model configuration; this may be due to duplication of the algorithm's function within a process group by a similar algorithm. In some model configurations, where more than one process algorithm had a low weight across catchments, more than one algorithm was replaced in the subsequent configuration for expediency and reduction of total model evaluations, as each model configuration required 2.4 million model evaluations in this experimental setup.

The model development process involved (a) the adjustment of selected process equations (Section 4.2.2), (b) addition of new blended process groups (Section 4.2.2), (c) testing of the so-called conglomerate model, (d) changes to the non-blended part of the model structure (Section 4.2.2), and (d) the reduction of complexity by removing process equations with marginal contribution to performance (Section 4.2.2). These specific developments are discussed in the following subsections. The specific algorithms included in process groups for each blended model configuration is provided in **Table7.4**.

Adjustment of selected process equations

Within each blended process group exist two or more process options. Iterating on various combinations and numbers of process options from the ones available within Raven is part of the model development process to empirically find options within a

process group where the overall model performance is improved, and all of the options within the process group are selected as preferable (indicated with calibrated process weights) in at least some of the tested catchments. An initial modification to the original blended model is the selection of new process options within each blended group based on the expert consideration of the authors that would perhaps improve model robustness and overall performance. The options adjusted by expert opinion may be viewed for model configuration 3 in **Table 7.4**. In cases where process options of a similar form or function exist (i.e., mathematical formulation, dependent state variables, etc.) within the same process group, one of the options may be replaced and this new configuration tested empirically to determine if a functionally different process option enhances the blended model configuration. The original blended model configuration selected options that worked well based on variations from the HMETS model (*Martel et al.*, 2017) rather than options that would fulfill the goals of this study.

Introducing blended forcing groups

The potential melt (POTMELT) and potential evapotranspiration (PET) estimators were both revised from their original (single) algorithm to blend two or more algorithms. Potential melt (mm/d) is defined as the maximum uncorrected snowmelt rate if snow is present, a surrogate for energy availability at the snow or land surface. Blending forcings in addition to hydrologic fluxes is a recent feature implemented in Raven and a contribution of this work, as this option was not available when developing the

initial version of the blended model. The potential melt and potential evapotranspiration forcings are estimated in Raven using precipitation and temperature data. The blending of potential melt and potential evapotranspiration follows the same form as other blended groups, i.e., as a weighted average of two or more process options. As long as algorithms are used to provide modelled estimates of the same units (e.g., PET and POTMELT both in mm/d), then these estimates may be combined as part of the same blended forcing group. These two forcings (PET and POTMELT) were selected for blending in this study because 1) they were considered the most influential estimated forcings within the hydrologic model, and 2) they had many options in Raven to experiment with. However, the same approach could be applied to other forcings in future versions, such as radiation estimation.

Conglomerate model

A so-called conglomerate model (model configuration 24) is tested and calibrated in the 12 MOPEX catchments with the same calibration budget of 10,000 runs. This conglomerate model is special in that it includes all process groups and process options tested in prior model configurations, includes 4 or more options in each process group with a total of 79 parameters. The second-largest model in terms of number of parameters has 59 parameters (model configuration 10). This model configuration is constructed to test whether the maximization of the number of process options included in the blended model configuration is advantageous relative to more carefully

curated model configurations; the process algorithms included within each blended process group are summarized in **Table 7.4**.

Non-blended structural changes

In addition to iterating on the blended process groups, non-blended changes in the model structure are tested in the model configurations. Model configuration 3 adds depression storage to the model, allowing for water to be stored on the landscape in depression storage and seep into soil storage. Later model configurations (34-37) add a set of canopy processes, including canopy interception, storage, drip and evaporation; the selection of non-blended (i.e. singular) process options for these are iterated on with different model configuration. The leaf area index (LAI) ratio is also allowed to vary seasonally in these model configurations (34-37) by introducing calibration parameters that modify the seasonal LAI input to the Raven model, allowing for either seasonal peaks in LAI during July and August or a flatter LAI ratio year-round as a function of new calibration parameters. Finally, processes are introduced in model configuration 34 that allowed the upward movement of water between soil layers, including from the third soil layer that was previously treated as a sink for deep groundwater. These were added as non-blended processes because it was considered to be of limited value added through blending due to the similarity of and/or limited number of algorithms for each type of process, though this may be tested in future studies especially as the process library in Raven grows for these processes.

Reduction in complexity

Attempts are made between the model configuration iterations to reduce the complexity (i.e., the number of process groups, process algorithms in each group, and the total number of parameters including weight-generating parameters) of a given blended model configuration by reducing the number of process options in one or more process group. While checks are made in the model methodology for overfitting (see Section 4.2.3), instances where the model complexity can be reduced by removing one or more options (and parameters) from the model without a reduction in performance in both calibration or validation across the 12 tested MOPEX catchments is generally regarded as a benefit. A guiding principle here is to ensure that any added complexity is improving the model performance, since it is later known from the conglomerate model test described in **Section 4.2.2** that including all possible options without such checks leads to a reduced model performance. This strategy of reducing model complexity was employed throughout the model development stages where possible.

4.2.3 Model overfitting

One of the assumptions made and checked in this study is that increasing the complexity in a blended model configuration is not a concern as far as model performance is considered. With a fixed calibration budget, it would be expected that at some point, increasing model complexity will result in a lack of convergence in calibration.

In order to more formally test this assumption, the model performance in validation for each blended model configuration (except for the conglomerate model) is checked against the number of parameters in each model configuration in a simple regression exercise. This is done across all model configurations and MOPEX catchments, and repeated within each of the 12 MOPEX catchments individually. In each case, a regression slope significantly less than zero would indicate a negative relationship between model performance in validation and the number of parameters, suggesting that more model parameters does reduce validation performance and model overfitting is an issue. This analysis is repeated using the validation gap (the change in performance between calibration and validation, defined in **Section 4.2.4**) instead of the validation KGE for both pooled results and individual MOPEX catchments. When the validation gap is used, a statistically significant positive slope would indicate that the discrepancy between the calibrated KGE score and its associated validation KGE score is increasing as more parameters are added.

4.2.4 Selection of a preferred blended model configuration

Following the development of over 30 model configurations, a single model configuration needs to be selected from the set of plausible candidates. While the KGE metric averaged over multiple independent trials and catchments is used to guide the development of model configurations, minute discrepancies in this metric may not be the most informative when picking from similarly performing model candidates (i.e., a mean

KGE of 0.865 may not be a better candidate model than one with a mean KGE of 0.864).

To aid in the evaluation of these numerous model configurations, additional metrics are used. The first is the difference between the maximum calibration KGE and the median calibration KGE for a given model configuration and catchment, then averaged across each of the 12 catchments. This metric is termed 'calibration consistency', with an ideal value of zero indicating complete consistency between the maximum and median calibration performance. Second is the maximum calibration KGE score minus the validation KGE score of the same trial, again averaged across 12 catchments. This is termed 'validation gap', and represents the drop in performance from calibration to validation. An ideal value would realistically be zero, indicating no drop in performance between calibration and validation, though the validation performance could also theoretically be greater, resulting in negative values. Third, the maximum calibration KGE performance, averaged across 12 catchments is used. Finally, the averaged validation KGE of 20 independent trials within a given catchment, then averaged across catchments is used. These last two metrics are simpler performance metrics relating to calibration and validation performance, respectively.

These four metrics (calibration consistency, validation gap, max calibration KGE, and mean validation KGE) are plotted in a set of two plots, which are used to discriminate between multiple high-performing model configurations. The pareto principle is applied in each case to determine which models outperform others on the plotted met-

rics, where a model that is better on both plotted axes can be considered to 'dominate' the performance of another model. The non-dominated models in these plots, i.e. the models that are better than any other model on at least one of the two plotted metrics, become part of the reduced set of candidate model configurations to be selected. The model configuration selected in this way becomes a new version of the blended model, referred to as version 2 of the blended model herein.

4.2.5 Validation of the selected blended model

The original blended model introduced by *Mai et al.* (2020) (model configuration 0 in this study) and the new blended model selected from plausible model configurations in this study (model configuration 36, now version 2) are evaluated with an independent set of 24 additional catchments, which are randomly selected from a subset of the HYSETS catchments (*Arsenault et al.*, 2020). This is considered a form of spatial validation, as these catchments were not used in the development or selection of the new blended model version 2. These 24 independent catchments were selected randomly from the set of catchments calibrated by *Mai et al.* (2022b), with the criteria that they were 1) also located in the United States, 2) had a total catchment area within range of the MOPEX catchments, and 3) had streamflow observations available within the calibration and validation periods selected previously for this study. In addition, the model performance results are evaluated for their statistical significance using the Mann-Whitney U test, a non-parametric test of the central tendency of two sets of

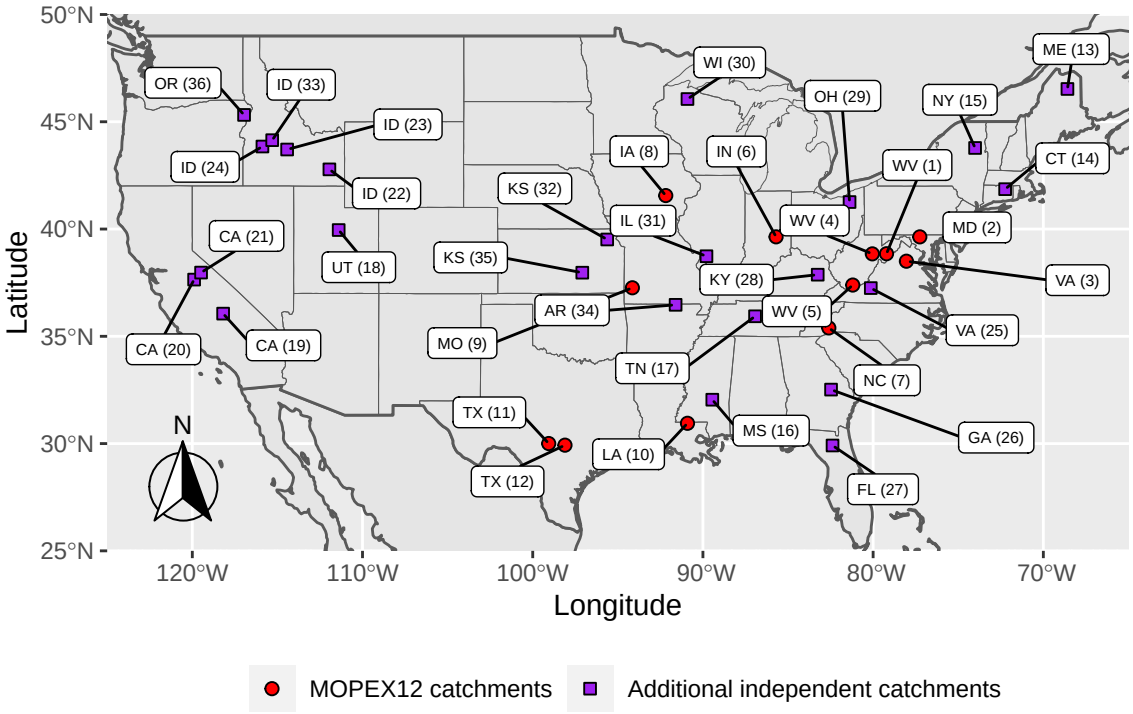


Figure 4.1 Map of the 12 MOPEX catchments and 24 additional independent catchments located in the continental United States. Captions indicate the state abbreviation and catchment index (i.e., 1 to 36).

sample data (*Hollander and Wolfe*, 1973). The map of the 12 MOPEX and the 24 independent catchments is provided in **Figure 4.1**. Additional information on the MOPEX and independent catchments is provided in **Table 7.3**, and characteristics of the 12 MOPEX catchments can also be found in *Duan et al.* (2006).

The evaluation of the two model versions (initial and final) is done by calibrating both model versions with the same setup as the original experiment, i.e., using DDS with a KGE objective function, the same warm-up and evaluation periods, and with

20 independent trials of each calibration in each of the 24 additional independent catchments. These results can then be viewed to compare the two model versions directly in terms of calibration and validation performance.

4.3 Results and Discussion

The following results and discussion sections are organized as follows. **Section 4.3.1** presents a summary of the developments in each blended model configuration and initial performance results. **Section 4.3.2** presents the statistical analysis undertaken to assess model overfitting. **Section 4.3.3** presents the selection process to determine version 2 of the blended model from the set of blended model configurations. Finally, **Section 4.3.4** describes the process to validate the blended model version 2 with a set of additional independent catchments.

4.3.1 Blended model version development and performance evaluation

The model development began with the initial blended model version, as published in *Mai et al.* (2020), which is referred here to as model configuration 0. The key developments by model configuration are summarized below, and are discussed in more detail in this section. Model configurations not listed below (25-28, 31-32) were used for the purposes of testing different weight-generation schemes in calibration (i.e. alternatives

to the weight-generating parameter setup described in **Section 4.2.1**), and did not contribute to an improved blended model version, and are therefore omitted from the results. A complete listing of the process options in each model configuration is provided in **Table 7.4**. Further descriptions of the process options in Raven may found in the Raven User's Manual v3.8 (*Raven Development Team*, 2024).

1. Model configuration 0: original model version from *Mai et al.* (2020);
2. Model configuration 1: model configuration 0 but with PET estimation from provided data (PET_DATA) rather than PET_OUDIN;
3. Model configuration 2: model configuration 1 but with an update to blended process options in process groups AET, baseflow, infiltration, quickflow, and snowbalance based on expert consideration;
4. Model configuration 3-5: sequential addition of depression storage & seepage to model configuration 2, followed by adjustments to other process options;
5. Model configuration 6-23: introduction of blended forcings process groups PET and potential melt, followed by iteration of process options;
6. Model configuration 24: testing of the conglomerate model configuration;
7. Model configuration 29-30, 33: experiments in reducing model complexity, particularly in the infiltration and snowbalance process groups;

8. Model configuration 34-37: addition of non-blended processes (including canopy processes), and additional complexity reduction experiments in the baseflow, potential melt, and quickflow process groups.

As expected with the nature of this type of empirical model development process, the calibration performances improve with successive model configurations, particularly in the initial configurations. The mean calibration KGE across catchments, i.e. the average of all 20 independent trials and 12 catchments, improved from 0.836 in the original blended version (model configuration 0) to 0.896 in model configuration 36, and the mean temporal validation KGE also improved from 0.728 to 0.799. The conglomerate model calibration performance was generally reduced relative to other model configurations (median calibration KGE reduced from 0.902 in model configuration 23 to 0.876 in the conglomerate model configuration 24). This reduction indicates convergence issues with the conglomerate model under the fixed calibration budget, and at a minimum suggests that the naïve inclusion of all (or many) plausible process options for limited calibration budgets may not be an optimal strategy. The performance of select model runs to summarize the changes with each stage of development are shown in **Figure 4.2**. The distribution of calibration performance results for select model configurations are provided in **Figure 4.3**, with validation performance of the maximum calibration performance trial plotted as a point when it falls within the plotting bounds.

The results of **Figure 4.3** show how the distribution in calibration performance

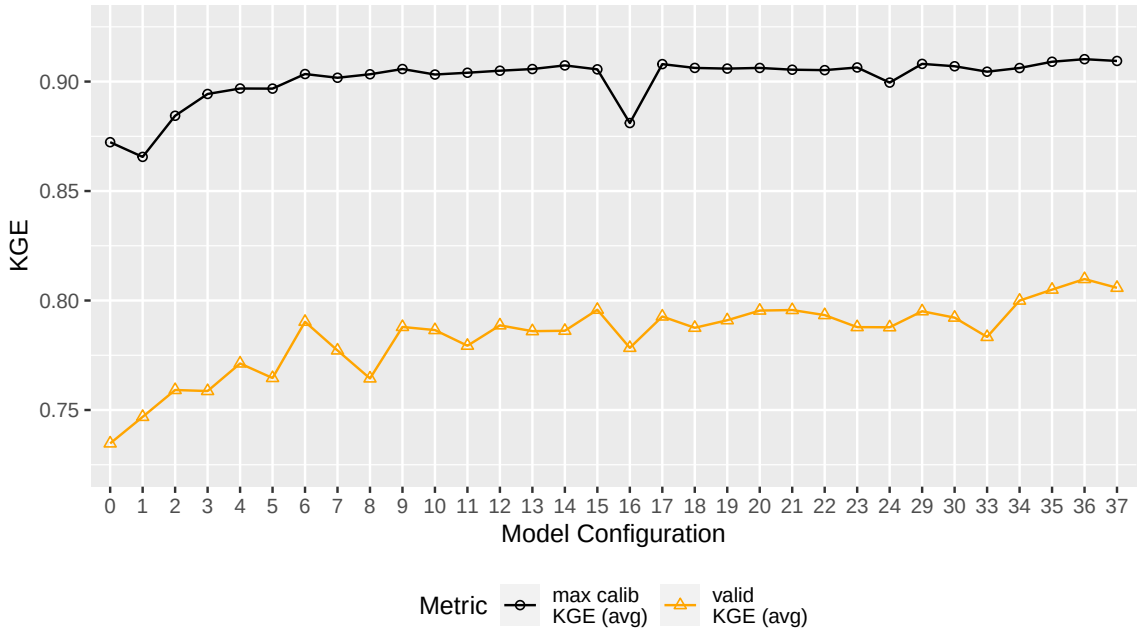


Figure 4.2 Model performance in calibration and validation by model configuration. The calibration KGE is the maximum KGE across 20 independent trials, and validation KGE is the associated KGE in the validation period of that maximum trial, both averaged across 12 MOPEX catchments.

between model configurations changes, where the initial model configuration 0 has relatively wide bounds in the box plots, indicating a wide range of likely outcomes in calibration, which holds true across most of the 12 MOPEX catchments. Similar results can be seen for model configuration 24 (the conglomerate model), as well some of the earlier model configurations. Overall, the width of the boxplots decreases with subsequent model configurations, indicating that the variability of calibration performance is decreased with iterative testing of model configurations. The results also show some interesting differences between the catchments, including the validation gap (difference in the calibration and validation scores), the calibration consistency (width of the calibration performance distributions), and the overall model performance between catchments. For example, the calibration consistency is generally much greater in some

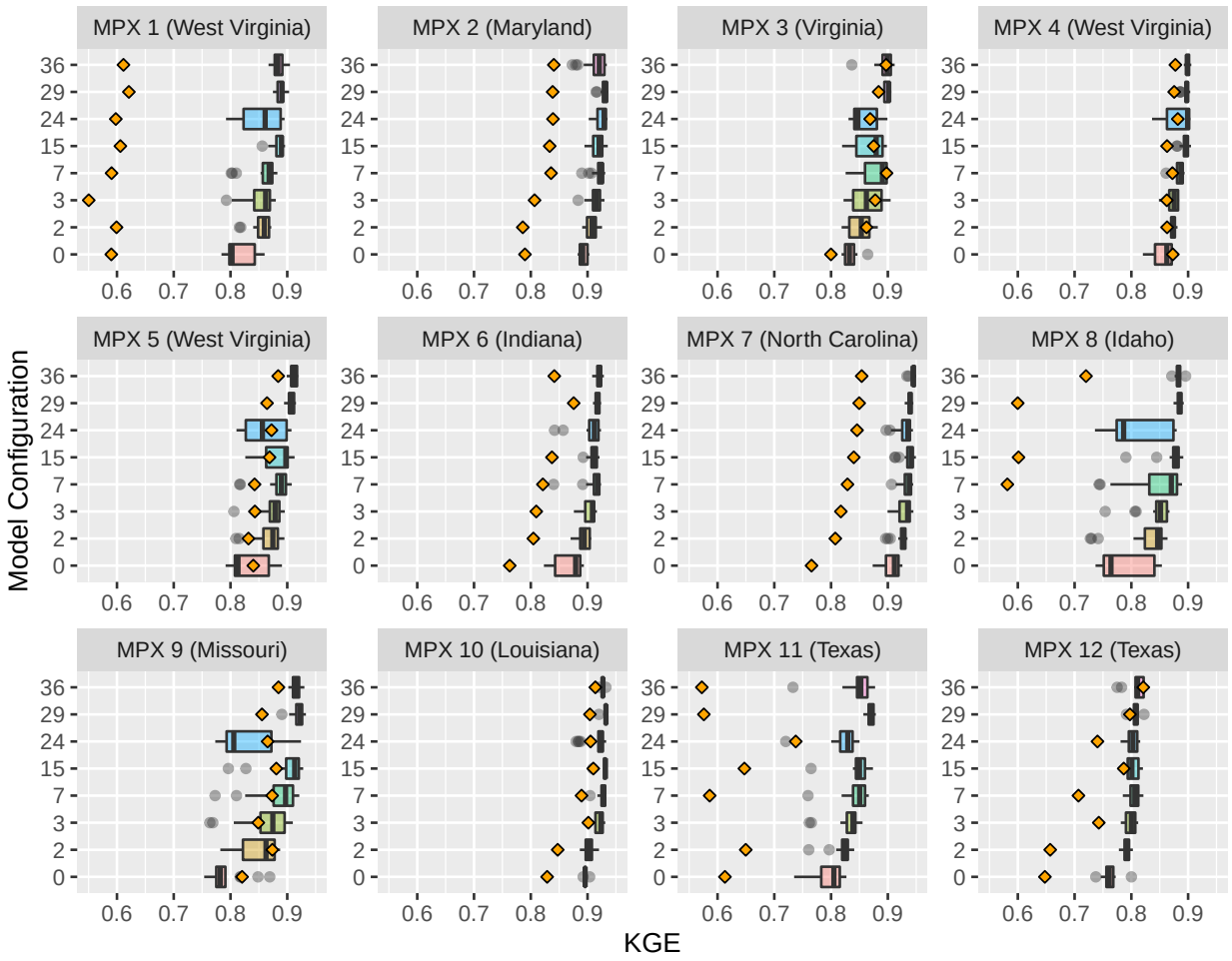


Figure 4.3 Calibration performance box plots, generated from 20 independent trials of the calibration for each model configuration and MOPEX catchment, shown for select model configurations. The orange diamonds indicate the validation performance of trial with the maximum calibration performance within the 20 independent trials (when it falls within plotting bounds).

catchments (e.g., 2 and 10) than in others (e.g., 3 and 9). As another example, the validation in MOPEX catchments 1 and 8 generally lag the calibration by a wider margin than other catchments. A series of linear regressions of performance metrics from model configuration 36 onto catchment characteristics (including the ones in **Table 7.3**) for all 36 catchments were run to look for possible explanations of the observed patterns. It was found that annual average precipitation and streamflow had statistically significant positive correlations with calibration and validation performance, and significant negative correlations with the validation gap. In simpler terms, catchments that had more precipitation and had more streamflow seemed to correlate to better performance in both calibration and validation scores, and had a smaller difference in the calibration and validation scores than catchments that received less precipitation and has less streamflow. No significant correlations were found to explain the calibration consistency in catchments within the results for model configuration 36. While no formal test was done to examine calibration consistency changes with model configuration, the calibration consistency generally seems to improve with subsequent model configurations. It is noted that model configuration 24 is the conglomerate model, which had substantially more parameters than other models and also a much larger variance in calibration and validation performance scores.

Additional metrics examining the differences between calibration and validation period conditions (precipitation and observed streamflow) were also extracted and regressed to try to explain the validation gap between catchments, though no statistically

significant correlation was found. In catchment 1 this difference in likely explained by the presence of a much larger flow event in the validation period than in the calibration period, making it difficult for the model to capture the magnitude of the unobserved event in validation. However, this was specifically high in catchment 1, and no general explanation was found for the variation based on calibration and validation period metrics between catchments otherwise.

4.3.2 Model overfitting

In order to assess overfitting in the blended model configurations, a series of linear regressions of validation performance against the number of model parameters were performed in order to look for statistically significant relationships. A linear regression was first applied to all model configurations and MOPEX catchments (save the conglomerate model, configuration 24), and second to individual MOPEX catchments. The conglomerate model was not included as it was a known case where the model was overfit from the performance results, and the goal of this analysis was to detect this for the other model configurations where it was assumed that overfitting was not occurring. The overall regression for all model configurations and catchments is shown in **Figure 4.4**, and the individual regressions are summarized in **Table 7.7**.

Figure 4.4 shows that across all catchments and model configurations there is no evidence of a linear relationship between validation KGE and the number of calibrated parameters, with a slope that is approximately 0.00032. When the regression is ap-

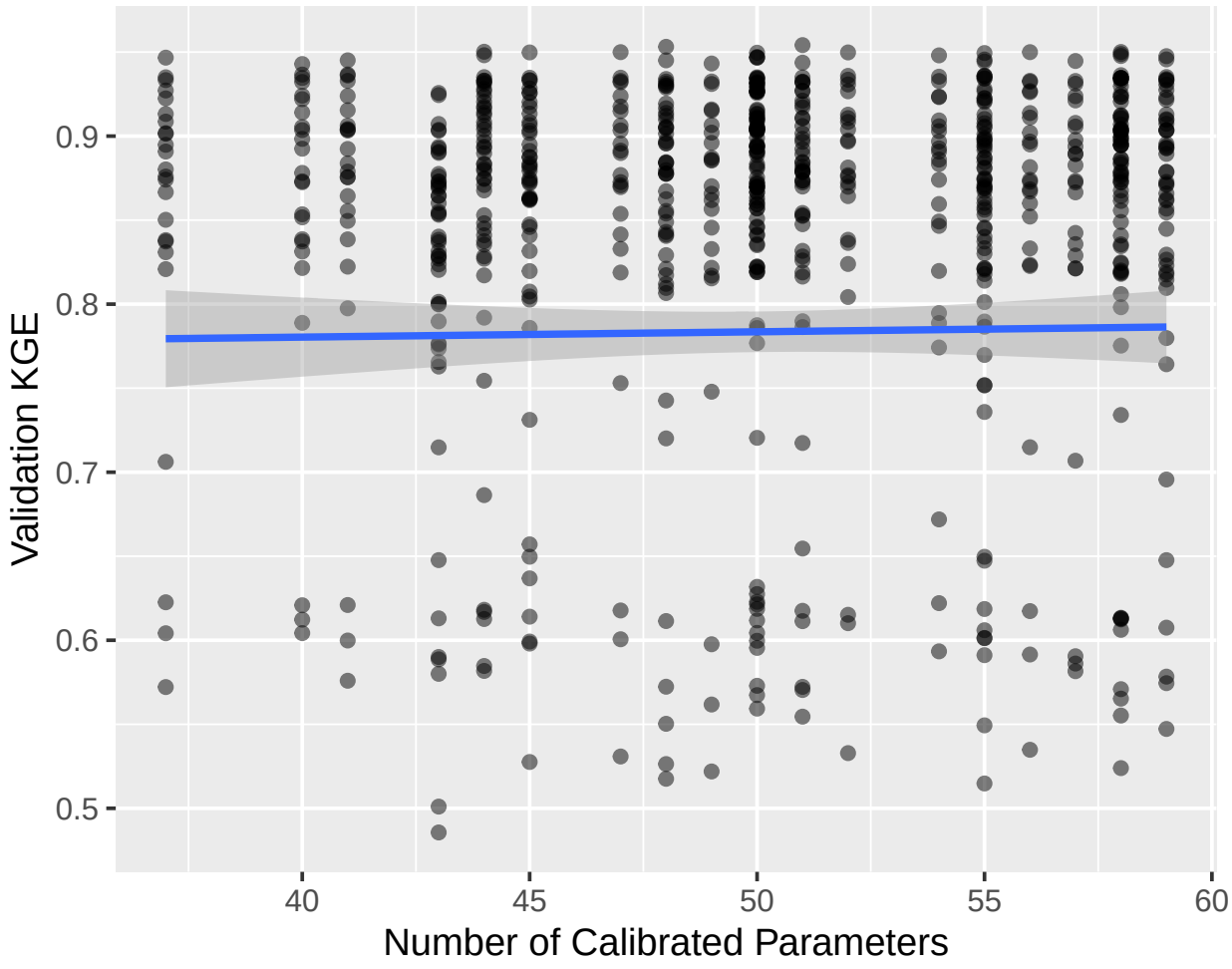


Figure 4.4 Validation KGE on the y-axis (obtained from the maximum calibration KGE trial) plotted against the number of calibrated model parameters on the x-axis for 31 model configurations and 12 MOPEX catchments. A regression line for this relationship is shown in blue as a nearly flat relationship (slope of $3.2E-4$ with a corresponding p-value of 0.757 for a null hypothesis of slope $\beta_1 = 0$), with a 95% confidence interval shown in transparent grey.

plied to all individual catchments, no catchment had significant evidence of a negative correlation which would indicate a decrease in validation KGE with an increase in the number of blended model parameters.

A similar analysis was repeated but using the validation gap (defined in **Section 4.2.4**) to detect decreases in relative validation performance with an increasing number of parameters. When this linear regression was applied across all model configurations and catchments, there was no evidence of a slope significantly different from zero (p-value of 0.96). This regression was then applied in all catchments individually, and these results are summarized in **Table 7.8**. These results show that only one catchment had a positive slope that was significantly different from zero (catchment 6 in Indiana, p-value of 0.081). This indicates that there is an increase in the validation gap at the 10% significance level, i.e. greater discrepancy between the calibration and validation KGE, with an increasing number of parameters. However, the p-value shows this finding is not incredibly strong, and the 11 other catchments had either no evidence of a slope significantly different from zero or had a negative slope (i.e., indicating that the discrepancy between validation and calibration decreases with additional parameters).

Overall, there is little evidence that additional calibrated parameters in the blended model configurations reduce the validation performance (absolute value or relative to calibration) in this experimental setup. However, as indicated by the conglomerate model, it seems likely that the validation KGE could degrade with an increase in the number of parameters at some point beyond 59 parameters, the maximum number

of parameters of model configurations in this regression. Similarly, a decrease in the calibration budget used here of 10 000 runs could also eventually affect convergence in calibration, likely degrading both the calibration and validation performance.

4.3.3 Selection of a preferred blended model configuration

The sequential improvement of the overall model performance is shown in **Figure 4.2**. However, a 'best' model would ideally be selected from this set of models, such that practitioners may have a single preferred blended model structure to work with. This task is made more difficult by the models which appear to have a similar performance, based on the mean calibration and validation KGE. This selection is therefore informed here by additional metrics (described in Section 4.2.4) which are plotted in **Figure 4.5**. In both plots A and B, the ideal point is set in the bottom left corner of the plot.

For **Figure 4.5A**, this ideal point is located at $[0, 0]$, and represents the point where calibration consistency is equal zero, meaning that the maximum calibration KGE is equal to the mean calibration KGE (i.e. all independent trials result in the same KGE score), and the validation gap is also equal to zero, meaning that the validation KGE is equal to the maximum calibration KGE. The validation KGE may theoretically be greater than the calibration score, but a model is generally not expected to perform better in validation than in calibration, and on average this assumption (calibration scores better than validation scores) holds in this experiment. The dashed line in this plot represents the Pareto front of non-dominated model configurations, i.e. model

configurations that are not clearly inferior in performance on these axes to another model configuration (namely, model configurations 36, 37, 34, 29, and 30). In **Figure 4.5B**, this ideal point is located at $[1, 1]$, where the mean of the maximum calibration KGE and the mean validation KGE are both equal to the maximum KGE value of 1.0.

In **Figure 4.5A**, a Pareto front of tradeoffs in the calibration consistency and the validation gap are generated, with model configurations 36, 37, 34, 29, and 30. It is shown that model configuration 0, the original blended model configuration, is dominated in this plot by all model configurations except two (model configuration 1 and 8), indicating the successive improvement of the models through the iterative process in this study.

In **Figure 4.5B**, only model configuration 36 is considered non-dominated, and model configuration 36 is also on the Pareto front in plot A. Again, it is shown that the original blended model (configuration 0) is dominated by every other model configuration, for which it is perhaps unsurprising given the empirical approach to modifying model configurations. The outcome here suggests that model configuration 36 is preferred over all of the options tested herein for use as a future baseline blended model. This outcome does not change if the metrics are rearranged between the two plots of **Figure 4.5**, though it is dependent on the setup of this study, such as the daily timestep, lumped model discretization, and available daily forcing data, and the preferred blended model configuration may change with different model decisions in other applications. The list of processes and process options for model configuration 36 are

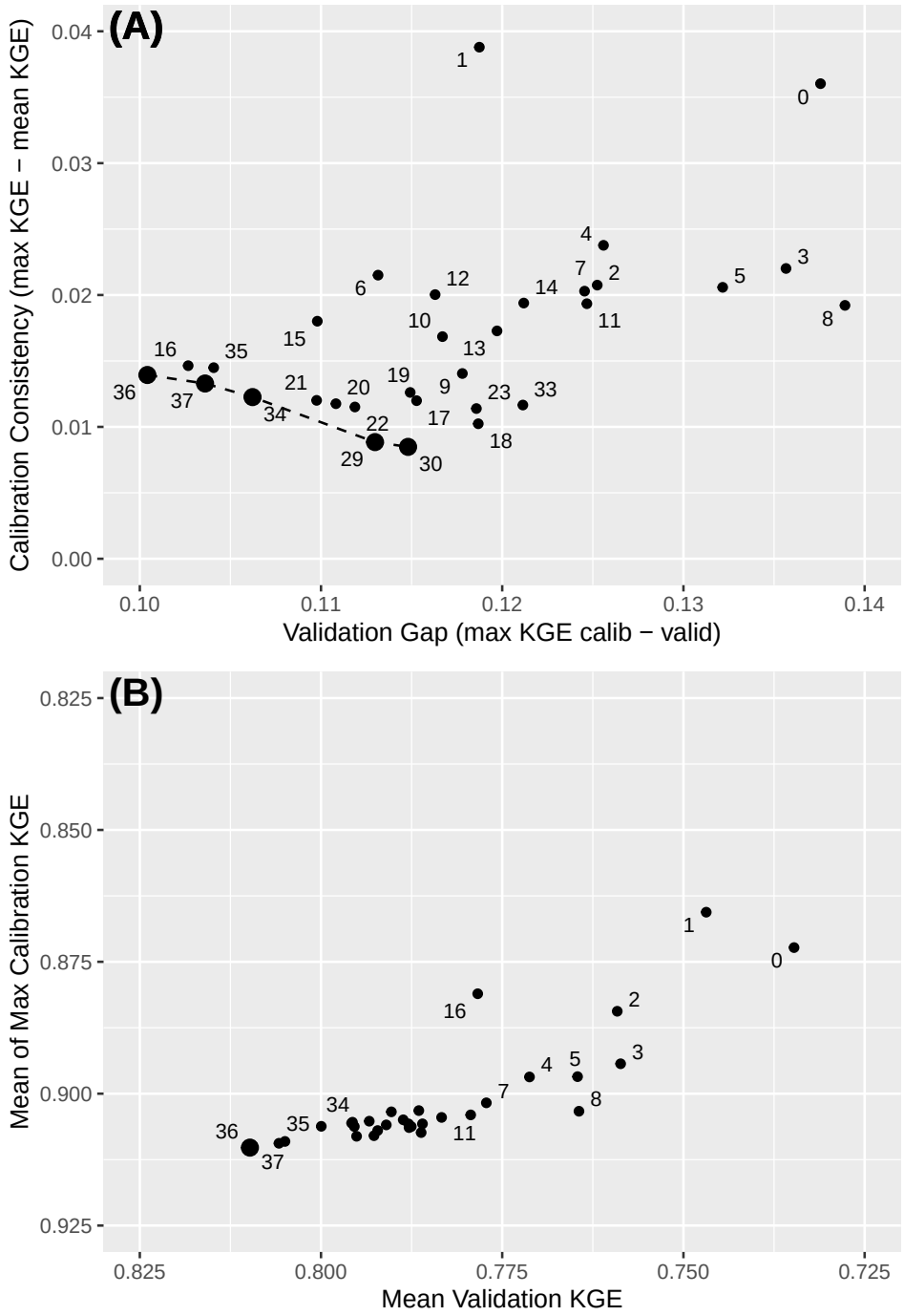


Figure 4.5 A) Model configurations plotted based on the calibration consistency on the y-axis against the validation gap on the x-axis, measured as the maximum KGE calibration performance minus its associated validation performance, averaged across 12 MOEPX catchments. The black dashed line shows the Pareto front of non-dominated model configurations (from left to right, 36, 37, 34, 29, and 30). B) The mean of the maximum calibration KGE performance plotted against the mean validation KGE performance, averaged across 12 MOPEX catchments. Pareto points are plotted with larger shapes than other data points.

provided in **Table 7.5**, and the parameter information for model configuration 36 is provided in **Table 7.6**. Raven model files are also available with the supporting data for model configuration 36 (see the Code and Data Availability section).

4.3.4 Validation of the selected blended model

The original and the newly selected blended model (model configuration 36) were calibrated on the independent set of 24 catchments as a form of spatial validation, since these new catchments were not used in the development or fine-tuning of either blended model configuration. The results for the two model versions in the independent catchments are summarized graphically in **Figure7.4**. The results are also analyzed using the Mann-Whitney U test to determine the statistical significance of differences in the median performance of both blended model versions. These results are provided in **Table 7.10**.

The statistical test results show that at the 5% significance level, there are 16/24 independent catchments where the calibration performance is significantly improved, and also 16/24 where the validation performance is significantly improved. There are 2 catchments where calibration performance is significantly greater in model configuration 0 than 36, and 3 where the same is true for validation performance. Overall, the selected model configuration 36 performs better (significant p-value with the median greater for model configuration 36) or statistically similar (insignificant p-value, i.e. greater than 0.05) to the original configuration in 22/24 (92%) of the independent

catchments for calibration performance, and 21/24 (88%) for validation performance. This suggests a marked improvement in model performance in both calibration and validation performance with the blended model version 2 over the initial version.

The mean differences for both calibration and validation for all 36 catchments are provided in **Table 7.10**. The average KGE improvement with the v2 model was 0.05 in calibration and 0.04 in validation, where this improvement exceeded 0.1 KGE for 4 of the catchments in calibration. A regression of the mean differences against hydrologic properties such as runoff ratio (**Table 7.3**) found a significant negative correlation (p-value ¡ 0.05) between the mean calibration difference and the runoff ratio (similarly a significant positive correlation with the evaporation index), indicating that the calibration score improvements were generally higher in drier catchments. This suggests that performance in drier catchments was an area of improvement in the original blended configuration.

4.4 Conclusions

The blended model configuration initially published in the literature has been shown in several studies to have a high performance when deployed. However, this initial configuration and overall model structure was fixed. In this study, we explore many blended model configurations and test their performance in the 12 MOPEX catchments located within the continental United States. In addition to testing different blended process groups, we introduce two blended forcing groups for potential melt and poten-

tial evapotranspiration, and we also introduce non-blended structural changes including processes for depression storage and canopy interception.

Of the more than 30 alternate model configurations explored, one is found to have a non-dominated performance upon examination of different calibration and validation metrics. This blended model configuration was then compared to the original version and evaluated against an additional 24 independent catchments within the continental United States. The blended v2 model was found to have statistically improved or indistinguishable performance in 92% of the independent catchments in calibration, and 88% in validation. The improvement in calibration performance was found to be statistically higher in drier catchments, suggesting this was an area of improvement in the original blended model configuration. We also tested for overfitting by performing a series of regression analyses of model validation performance (both absolute and relative to calibration performance) against the number of model parameters, and found no evidence of a decrease in validation performance with additional model parameters within the conditions of our calibration experiments.

This study provides several considerations for future development of blended model configurations, including approaches for evaluating process options when designing a blended model, new options for blended forcings, and strategies to reduce the complexity of blended models. This study also delivers an improved blended model as version 2, with a demonstrated increase in calibration and validation performance for catchments in the continental United States. This blended model version provides

enough flexibility to be robust across a range of various catchments without the need for adjusting its structure beyond what is done in model calibration. This blended structure should be of interest to practitioners, as the model structure is openly and freely available, and has demonstrated high performance 'out-of-the-box' in many cases without a need to undergo the iteration on model structural options to achieve a good fit in calibration for a particular catchment. The blended model version 2 may be used in future applications where high model performance is required, and may also be used in addressing scientific questions around the structural uncertainty and selection of processes in hydrologic models.

Chapter 5

Rapid Floodplain Mapping at Larger Scales

Statement of Contribution

The following chapter is a partially modified version of the following preprint article (*Chlumsky et al.*, 2024b):

Chlumsky, R., Craig, J. R., and Tolson, B. A. (2024): A reach-integrated hydraulic modelling approach for large-scale and real-time inundation mapping. Geosci. Model Development Discussions, 2024, 1-28, doi:https://gmd.copernicus.org/preprints/gmd-2024-184/.

This chapter presents a novel hydraulic modelling and mapping approach, the Geospatially Augmented Standard Step (GASS) method, which is designed for large

scales and suitable for real-time deployment. The method is implemented in the software package Blackbird and benchmarked against the HAND method and HEC-RAS 1D and 2D models. In this chapter, Robert Chlumsky wrote the Blackbird software package, designed the benchmarking experiment, gathered input data and prepared all models included in the experiment, performed the analysis of results and wrote the chapter. Dr. Craig and Dr. Tolson contributed to conceptual discussions in development of the Blackbird code, and provided their review and feedback on the chapter. Brian Peng contributed to the development of C++ versions of the Blackbird code to improve computational efficiency.

5.1 Introduction

Flooding is one of the most common and costly natural hazards, with expected trends showing an increase in both exposure and vulnerability to flooding globally (e.g., *Bhupsingh et al.*, 2022; *Aristizabal et al.*, 2023; *Steinhausen et al.*, 2022). As an example, the total residential flood risk in Canada is estimated to be $2.9 billion CAD per year by Public Safety Canada (*Bhupsingh et al.*, 2022), while the availability of flood maps in high-risk Canadian communities is ranked as poor (*Henstra et al.*, 2019). There is a need to mitigate flood risk for both the protection of human lives but also against financial losses. While there are different types of flooding (largely categorized as fluvial, pluvial, coastal, and minor system flooding including sewer backups), here we focus on addressing fluvial or riverine flooding (i.e., flooding that results from the overtopping of river banks as the river's capacity is exceeded).

Tools for mitigating flood risk and estimating flood extent come in different forms. For example, satellite imagery can be used to derive near real-time or rapid flood maps (e.g., *Tripathy and Malladi*, 2022; *Cian et al.*, 2018). These approaches are based on comparing satellite imagery during flood events to dry events, and delineating flooded areas based on the differences in imagery. The availability of social media and other online information has also been exploited through machine-learning approaches and data mining to help determine active flood locations (e.g., *Lin et al.*, 2023; *Li et al.*, 2018; *Fohringer et al.*, 2015). While these techniques may be useful in obtaining

near real-time conditions, they are not useful in prediction or simulation of future or possible flood impacts, as they rely on the collection of data in the period during and shortly after flooding occurs. In order to obtain simulations or forecasts of flooding, particularly under conditions not observed in the historical record, numerical models that can estimate future or hypothetical conditions are required.

Existing hydrodynamic models based on mass, energy and momentum conservation, such as HEC-RAS (*Hydrologic Engineering Center*, 2023), have been used extensively to develop flood maps. These typically involve a numerical solution to the shallow water equations, either in one dimension (1D) or two dimensions (2D). The results of 1D models can be imputed in space to create flood maps, though imputation is frequently done with an external program (such as HEC-GeoRAS or external GIS programs, (e.g., *Yang et al.*, 2006; *Md Ali et al.*, 2015; *Hashim et al.*, 2021)) and issues exist in the ease of generation and the interpretation of these results from 1D models with cross-sections. These include the prediction of wetted areas in locations with no hydraulic connection to the channel, caused by linear interpolation of water surface elevations (*Bates*, 2022). The accuracy of 2D models is generally considered much higher, with a number of studies confirming the benefit of 2D models over 1D in most circumstances (*Ghimire et al.*, 2022; *Jafarzadegan et al.*, 2023). In addition, 2D models do not require interpolation in the same way as 1D models since results are already spatially distributed. However, 2D models have a number of drawbacks, including high effort to configure and a high computational expense, and are therefore difficult

to implement at large scales. Recent research efforts have focused on improving the efficiency of 2D models, including use of simplified forms of the underlying shallow water equations, hybrid 1D-2D approaches, code optimization, and leveraging the use of high-performance computing (*Jafarzadegan et al.*, 2023). Other approaches have also attempted to use machine-learning tools to either emulate 2D model results (e.g., *Kabir et al.*, 2020; *Zhou et al.*, 2022), or to simulate inundation directly (e.g., *Shaeri Karimi et al.*, 2019; *Frame et al.*, 2024). Global flood maps have also been produced at an approximate resolution of 90m using the 2D LISFLOOD-FP model (*Sampson et al.*, 2015). In all of these cases, the high degree of effort and computational cost of running the hydrodynamic models remains a barrier in practice, particularly for large scale, high-resolution, and/or real-time prediction.

The computational demands of hydrodynamic models has at least partially driven the development of simplified methods for flood inundation and mapping (*Hamidi et al.*, 2023). Many simplified methods focus on use of high-resolution topographic data to estimate flood depths. The most basic version of this could be considered the planar method, where any raster pixel with an elevation less than a specified flood elevation is considered flooded (*Teng et al.*, 2015; *McGrath et al.*, 2018). Other examples of recent simplified methods include the rapid flood spreading method (*Lhomme et al.*, 2009b), AutoRoute (*Follum*, 2013), and the Height Above Nearest Drainage (HAND) model (*Rennó et al.*, 2008; *Nobre et al.*, 2011). The popular HAND approach and its derivatives have been used in many studies (e.g., *McGrath et al.*, 2018; *Scriven et al.*, 2021;

Chaudhuri et al., 2021; *Aristizabal et al.*, 2023) as a low complexity, low computational cost approach to estimating flood depths and inundation extents, which is also scalable to large domains. The HAND approach re-expresses an elevation raster in terms of relative elevation to the nearest drainage point, which can then be used to rapidly generate inundation maps with depths for each drainage feature provided. Approaches in the literature for generating these depth estimates typically deploy Manning's equation to develop synthetic rating curves that estimate depth from flow (e.g., *Afshari et al.*, 2018; *Scriven et al.*, 2021; *Chaudhuri et al.*, 2021; *Diehl et al.*, 2021), and relies on the calculation of hydraulic properties over a reach to estimate wetted perimeter and area in Manning's equation (*Zheng et al.*, 2018a,b). This type of approach, referred to here as the HAND-Manning method, has been shown to work reasonably well at emulating results of more sophisticated models in relatively simple cases, but struggles with the very common case where backwater effects influence results (*Aristizabal et al.*, 2023), since each domain or catchment area the method is applied to is computed independently. Hydraulic structures can also not be represented with this approach. In addition, these methods do not respect mass, energy or momentum conservation. Other studies have circumvented these particular issues by calculating depths at cross-sections using HEC-RAS, and then mapping these depths using HAND rather than interpolating results between cross-sections (*Li et al.*, 2023). However, this approach assumes that the information at cross-sections is representative of conditions along the entire reach without considering the reach data in hydraulic calculations, which may not be appropriate in all cases.

In this study, we introduce a novel approach that combines the inundation mapping capabilities of the HAND approach with a 1D steady-state hydraulic model capable of handling backwater effects. We also provide a variant of the HAND approach, the Dynamic HAND (DHAND), which improves upon the ability of HAND to capture landscape connections within the DEM. The HAND or DHAND approaches are used in estimating hydraulic properties for 1D calculations and in mapping the depth results, thus removing the need for cross-sections that are typically deployed in 1D hydrodynamic models. This overcomes the limitations of HAND-only approaches with no backwater effects or energy loss calculations, and also integrates the complete set of 3D information into the hydraulic model rather than relying on cross-sections to extract local data. The HAND-based method is used both in determining hydraulic properties for the hydraulic model (rather than from a cross-section) and in mapping the depth results in post-processing. These methods are combine within the Geospatially Augmented Standard Step (GASS) method, and implemented within the Blackbird software package.

The main objectives of this study are:

1. To present of a methodology that utilizes the reach-integrated approach and the HAND model to represent hydraulic properties, and combine it within a 1D hydraulic model;

2. To introduce a modified version of HAND that captures the landscape connectivity as a function of stage, and improves upon the ability of HAND to capture

landscape connections in complex terrains; and

3. To demonstrate of the ability of the Blackbird software (which implements the GASS method) to estimate channel depths by first verifying the code against standard cross-section based models, then benchmarking it against similar methods for hydraulic calculations and mapping to contrast performance.

5.2 Methods

5.2.1 Geospatially Augmented Standard Step (GASS) overview

The Geospatially Augmented Standard Step (GASS) method uses a combination of three main approaches:

1. a steady-state hydraulic model solved by the standard step method
2. a dynamic version of the HAND method (DHAND), which is used in assessing whether cells are inundated for a given calculated downstream depth in pre- and post-processing
3. calculation of reach-integrated hydraulic properties in domains divided by overland flow paths, which are calculated prior to simulation to maintain a fast model runtime

The use of a standard-step method has been deployed in hydraulic models such as HEC-RAS for decades (*Brunner*, 2022). The DHAND method is a novel approach

developed specifically herein, and similar strategies for reach-integration of hydraulic properties have been introduced in previous studies (*Zheng et al.*, 2018a,b). Here, we extend these reach-integration strategies to additional hydraulic properties that are required in computing depths with the standard step method, such as conveyance and velocity coefficients.

The data requirements of this methodology include a digital elevation model (DEM) which includes observed or estimated bathymetric, spatially distributed Manning's n roughness values, a set of main channel polylines, and flows defined at each channel reach. The resolution of the input raster layers is up to the user, though the resolution should be sufficient to capture the channels being modelled and model outputs will be at the same resolution. The roughness layer can be determined from land cover information and translated to roughness values using the guidance of *Chow* (1959) or similar roughness tables. The channel polylines may be determined through processing of the input DEM to ensure alignment with the DEM. The bathymetric information is often not available in regional studies, though techniques for estimating channel depth (e.g., *Wilkerson and Parker*, 2011; *Thayer*, 2017), and reflecting this in the DEM through methods such as stream-burning also exist in the literature (e.g., *Lindsay*, 2016). The key output from the model is a spatially-distributed raster file of flood depths for each vector of flow values provided at each streamnode.

The basic workflow for this method is:

1 Input data collection and pre-processing

1a) collect inputs including DEM and Manning's n roughness raster layers, which would be supplied or processed to the same resolution

1b) determine main channel locations with a supplied polyline dataset or through a flow accumulation exercise, which may include manual delineation or pruning of channels to ensure channels of interest are represented

1c) generate the flow accumulation, flow direction, and HAND or DHAND layers

1d) discretize the model by defining streamnodes (locations where depths will be computed, similar to cross-section locations in a fully 1D model), and compute reach-integrated hydraulic properties as a function of depth for each streamnode (this is done once as a pre-processing step)

2 Simulation and post-processing

2a) provide flows for each reach, set computational options including boundary conditions, and perform the calculation of depths at each streamnode using reach-integrated hydraulic properties with the standard step method

2b) post-process results by interpolating depths along the reaches, and using the HAND or DHAND raster(s) to determine the depth of flooding in each cell

3 (optional) Iteration

3a) iterate on options and parameters for calibration, such as adjustment of a roughness multiplier (which does not require reprocessing hydraulic properties)

1. collect inputs including DEM and Manning's n roughness raster layers, which

would be supplied or processed to the same resolution

2. determine main channel locations with a supplied polyline dataset or through a flow accumulation exercise, which may include manual delineation or pruning of channels to ensure channels of interest are represented

3. generate various geospatial layers, including the flow accumulation, flow direction, slope, and HAND or DHAND layers

4. discretize the model by defining streamnodes (locations where depths will be computed, similar to cross-section locations in a fully 1D model), and compute hydraulic properties as a function of depth for each streamnode (this is done as a pre-processing step)

5. provide flows for each reach, set computational options including boundary conditions, and perform the calculation of depths at each streamnode using the standard step method

6. post-process results by interpolating depths along the reaches, and using the HAND or DHAND raster(s) to determine the depth of flooding in each cell

7. (optionally) iterate on options and parameters for calibration, such as adjustment of a roughness multiplier (which does not require reprocessing hydraulic properties)

The flooded area for a given reach is shown schematically in **Figure 5.1**, and provides a representation of how the GASS method discretizes the spatial domain. The

figure depicts a) the view of an inundated reach in pre-processing, b) the plan view of an inundated reach area delineated by the topographic divide and the steepest path divide at streamnodes, and c) an inundated reach after the hydraulic model is used to compute depths at each streamnode, and depths are first interpolated between streamnodes and then mapped with a HAND-based method to generate the final inundation map. The terms in the diagram are described throughout this section.

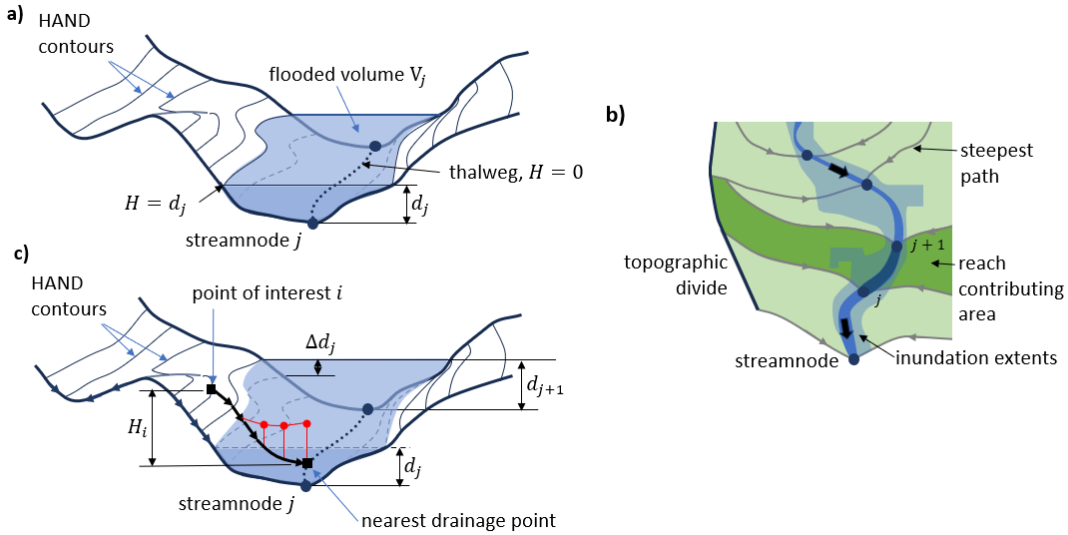


Figure 5.1 a) Schematic of the reach-integrated flooded area, b) plan view of the inundated area, and c) schematic of the inundated reach with interpolation of depth from streamnodes j and $j+1$. The water surface elevation is held constant along the red line in c), which is the vertical projection of the drainage path ending in the nearest drainage point.

5.2.2 Hydraulic modelling with the standard step method

In the traditional cross-section approach, the standard step method may be used to compute steady state depths with gradually varied flow at each streamnode. This method is identical to that used in HEC-RAS (*Brunner*, 2022).

The standard step method relies on solving Bernoulli's equation without the pressure head term for each pair of streamnodes:

$$z_j + d_j + \frac{\alpha_2 \bar{v}_j^2}{2g} = z_{j-1} + d_{j-1} + \frac{\alpha_1 \bar{v}_{j-1}^2}{2g} + \Delta h_j$$

where z is the channel invert, d is the channel depth, $\bar{v}$ is the average cross-sectional velocity, α is the velocity coefficient, g is gravitational acceleration, and h_e is the head loss between two streamnodes. The subscripts refer to the relative location of the streamnode, where j is a given streamnode and $j-1$ is the downstream streamnode. The program currently assumes a subcritical regime, applies a normal depth boundary condition at the downstream-most node ($j = 1$), and then proceeds upstream to solve each streamnode in turn.

The head loss between two streamnodes Δh_j is expressed as:

$$\Delta h_j = L_{e_j} \bar{S_{f_j}} + C_j \left(\frac{\alpha_j v_j^2}{2g} - \frac{\alpha_{j-1} v_{j-1}^2}{2g} \right)$$

where L_{e_j} is the effective stream length, S_{f_j} is the friction slope, and C_j is the expansion or contraction loss coefficient, depending on whether the flow area is larger or smaller than that of the downstream streamnode (A_j). The friction slope is computed at each streamnode using the relationship:

$$S_{f_j} = (Q_j / K_j)^2$$

where Q_j is the flow rate at the downstream streamnode and K_j is the total conveyance of the streamnode. The average friction slope $\bar{S}_{f_j}$ may be computed in a number of ways, where the default in Blackbird is a simple reach average between the downstream and upstream friction slopes. Additional details on the standard step method may be found in *Brunner* (2022).

While the focus in Blackbird is on the use of reach-integrated hydraulic properties (discussed in **Section 5.2.4**, the expression of hydraulic properties into equivalent form allows properties at streamnodes to be determined from either the traditional cross-section or from reach-integration, or a mix of both within the modelling domain, without a change in the standard step solution. This also allows Blackbird to theoretically include any feature of 1D cross-section based models as well, such as hydraulic structures. This removes a common limitation in large scale flood mapping applications based on geospatial methods such as the HAND-Manning approach (e.g., *McGrath et al.*, 2018; *Chaudhuri et al.*, 2021), which are unable to handle backwater effects or hydraulic structures.

5.2.3 Dynamic Height Above Nearest Drainage (DHAND)

In previous studies (e.g., *Afshari et al.*, 2018; *Scriven et al.*, 2021; *Chaudhuri et al.*, 2021; *Diehl et al.*, 2021), the HAND method is used to determine whether a cell is flooded based on its HAND value (H_i) relative to the depth (assumed uniform) in the reach (d_j). A cell is flooded with a depth of $d_j - H_i$ if $d_j > H_i$. One issue with the

processing required to produce a HAND raster is the need to condition the DEM using a breach or fill algorithm prior to computing the HAND raster (e.g., *Scriven et al.*, 2021). This is a common approach in watershed delineation, but in determining flow pathways at a high resolution for flood mapping, this can either create artificial flow pathways that do not exist in the terrain if a breach conditioning algorithm is used, or remove storage from the landscape if the fill algorithm is used. This is a known issue with the processing behind the HAND method (*Aristizabal et al.*, 2023). In any case, the generation of the HAND raster itself requires manipulating the DEM to determine flow pathways.

The Dynamic Height Above Nearest Drainage (DHAND) is presented here as an alternative to HAND that minimizes the artefacts of terrain alterations through conditioning, and is meant to preserve the flow connections in the DEM. This can be of particular importance when natural berms or engineered dykes exist in the landscape. The DHAND algorithm computes a series of raster layers with HAND values as a function of depth, where the number of depth intervals (N_k) and the depth interval between layers (Δd_k) are specified by the user. The calculation of DHAND starts with the initial HAND estimate determined from regular conditioning, which can be considered the DHAND layer for a depth of zero ($d_k = 0$). Then, for each specified depth $d_k = k \cdot \Delta d_k$,

1. a new DEM raster is constructed for the given layer k, where cells with a HAND value H_i of less than d_k in layer d_{k-1} are conditioned with a breaching algorithm,

and all other cells are conditioned via a fill algorithm

2. the HAND layer for depth d_k is then computed in the standard way, subtracting the elevation of the cell where a given cell drains to ($z_{i_{min}}$) from each cell's elevation (z_i) to compute the DHAND layer, i.e. $H_i = z_i - z_{i_{min}}$

This use of conditional breaching and filling of the DEM allows the landscape connections to be largely preserved, and particularly ensures that cells at lower elevations do not artificially become inundated for a given depth because of the conditioning algorithm. The accuracy of this method is improved with higher resolution depths applied in DHAND, at the cost of additional pre-processing and storage requirements, as each depth produces a new raster DHAND layer.

This is shown graphically in **Figure 5.2**.

5.2.4 Reach-integrated hydraulic properties

The depth in each cell (d_i) is determined by comparing the 1D reach-integrated depth d_j to the DHAND value of the raster cell (H_i). This assumes a flat raster cell where slope and aspect are neglected. The depth in each cell is calculated as $d_i = \max(d_j - H_i, 0)$.

The flooded area (A_i) is a binary function, defined as:

$$A_i = \begin{cases} 0 & \text{if } d_i = 0 \\ A_i & \text{if } d_i > 0 \end{cases} \tag{5.1}$$

where a is the square raster cell dimension. Similarly, the flooded volume V_i is computed as $V_i = A_i \cdot d_i$ the ith cell is inundated, and zero otherwise. The effect of slope being

incorporated into these equations was tested and found to have a negligible impact on results, and thus the simpler equations are adopted in calculating area and volume for raster cells.

In a traditional cross-section based 1D model, we compute conveyance in each section of the cross-sectional flow area as:

$$K_b = \frac{1}{n_b} A_b R_b^{2/3}$$

where K_b is the conveyance of the bth section, n_b is the Manning's roughness, A_b is the flow area, and R_b is the hydraulic radius, computed as A_b/P_b, where P_b is the wetted perimeter.

To compute the total conveyance in a cross-sectional model, we can sum the conveyance in each section:

$$K_j = \sum_{b=1}^{N_b} K_b$$

where K_j is the total conveyance of the cross-section, and N_b is the number of sections (or slices) within the cross-section. The number of sections or slices is typically either based on the number of ordinates (i.e., each pair of x, z coordinates constitute a new section) or the sections are grouped based on whether the roughness value changes, which results in fewer sections.

In the reach-integrated approach we consider each raster cell as a 'section' and substitute N_b sections for N_i cells. We further substitute the cross-sectional flow area for total cell flooded volume V_i, and the cross-sectional wetted perimeter for the total

cell flooded area A_i. The term is then divided by the effective length of the reach, L_{e_j}, which creates a term that is comparable to that traditionally computed for a cross-section. This approach is similar to that of *Zheng et al.* (2018a), though we use the effective length rather than reach length, as the effective reach length is more representative of the flow path as a function of depth than a static reach length. This is written as:

$$K_j = \frac{1}{L_{e_j}} \sum_{i=1}^{N_i} K_i = \frac{1}{L_{e_j}} \sum_{i=1}^{N} \frac{1}{n_i} V_i R_i^{2/3}$$

where in the reach-integrated set of properties, $R_i = V_i/A_i$, which is analogous to A/P in the original Manning's equation and equivalent in a uniform channel cross-section. We can similarly compute reach-integrated area and wetted perimeter, and compute the flow area of the jth streamnode (A_j) as:

$$A_j = \frac{1}{L_{e_j}} \sum_{i=1}^{N_i} V_i$$

and the wetted perimeter (P_j) of the jth streamnode as:

$$P_j = \frac{1}{L_{e_j}} \sum_{i=1}^{N_i} A_i$$

The effective reach length of the channel is intended to capture the distance travelled by water from one point to another, and is then used to determine energy losses incurred by that travel. In traditional 1D models, this is resolved by inputting the main, left, and right channel distance manually ($N_b = 3$), and often by using a flow-weighted arithmetic mean to determine the effective length for energy loss. In the cross-section

based implementation of our model, we use a similar conveyance-weighted approach to determine effective reach length:

$$L_{e_j} = \frac{1}{K_j} \sum_{b=1}^{N_b} L_b \cdot K_b$$

where $N_b = 3$ and represents the left, right and main channel sections.

In the reach-integrated implementation, we aim to determine the effective reach length automatically based on the geospatial data available without the need to prescribe the channel lengths for three locations (left, main, right), as is traditionally done in the cross-section approach. Instead, we compute a reach-length raster input, which determines the mean travel path length within the reach for a given stage. Conceptually, this reach-length raster represents the average path length that a droplet of water, if travelling through the reach at the given location and elevation of the raster cell it is located on, would travel from the upstream boundary of the reach to the downstream boundary of the reach. Within the channel this is determined from the channel vector polyline, and elsewhere this is determined from contour lines derived from the HAND raster. The path lengths mapped as a set of polylines are then interpolated to the DEM resolution to generate the reach length raster.

The effective reach length (L_{e_j}) of the jth streamnode for a specific depth is computed as:

$$L_{e_j} = \frac{1}{\sum_{i=1}^{N_i} K_i} \sum_{i=1}^{N_{cells}} L_i * K_i$$

where L_i is the estimated reach length in the ith cell (determined from the reach length

raster), and K_i is the cell conveyance.

Two other terms that need to be computed in order to facilitate steady-state hydraulic calculations include the alpha term (α), which is a correction to the average velocity, and the composite Manning's n (n_c).

The alpha term (α) accounts for the underestimation of mean velocity when computed from the sum of individual components. This value is generally between 1.0 and 4.7 (*Hulsing et al.*, 1966). In a cross-section approach, the alpha term has been computed as (*Brunner*, 2022):

$$\alpha_j = \frac{A_j^3}{K_j^2} \left(\sum_{b=1}^{N_b} \frac{K_b^3}{A_b^2} \right)$$

We compute α_j in the same manner with the catchment approach, noting that the the conveyance K_b is now the conveyance of an entire cell (K_i), and the cross-sectional area A_b becomes the total flooded volume of the cell V_i. This can be written as follows:

$$\alpha = \frac{V_j^{f3}}{K_j^2} \left(\sum_{i=1}^{N_i} \frac{K_i^3}{V_i^{f2}} \right)$$

The composite Manning's n in the cross-sectional approach (n_c) can be computed using one of many equations (see *Chow* (1959), equations 6-17 through 6-19), including the equal force equation, equal velocity, weighted average conveyance, weighted average area, and others. No single method has been deemed the most appropriate in the literature. Here, we again substitute the cross-sectional area with the flooded volume and substitute the cross-sectional wetted perimeter with the flooded cell area. The default composite Manning's n estimation approach in Blackbird is the equal velocity

method, which can be written in reach-integrated form as:

$$n_c = \left(\frac{\sum_{i=1}^{N_i} A_i n_i^{1.5}}{A_j} \right)^{2/3}$$

5.2.5 Post-processing and flood map generation

Following the execution of the hydraulic model calculations with a standard step approach, the mapping of hydraulic results at each streamnode is performed to generate a flood map of depths and/or inundation extents. The typical approach with HAND is to apply the reach-integrated depths d_j within each reach, and compute depths in each cell as $d_i = \max(d_j - H_i, 0)$.

In Blackbird, the thalweg depths are first interpolated along the reach, reducing the occurrence of sharp differences in depths within the flood map. This creates a series of interpolated depths $\hat{d}_j$ along the reach, where $d_j <= \hat{d}_j <= d_{j+1}$. The interpolated depths $\hat{d}_j$ may optionally be corrected with the term $(\hat{z}_t - z_t) * C_t$ to take into account the variability of the DEM elevations at the thalweg, such that the final water surface elevation remains a smooth linearly interpolated grade between streamnodes j and $j + 1$, rather than a water surface that reflects the variability of bed elevation in the channel. The interpolated depths $\hat{d}_j$ are computed as:

$$\hat{d}_j = d_j + (d_{j+1} - d_j) * \frac{\hat{L}_j}{L_{e_j}} + (\hat{z}_t - z_t)$$

where $\hat{L}_j$ is the length of the thalweg from streamnode j to the nearest drainage point associated with $\hat{d}_j$, $\hat{z}_t$ is the linearly interpolated channel bed elevation, and z_t

is the actual bed elevation (i.e., thalweg elevation). Where the depths are corrected by thalweg elevation in post-processing, the same correction would be applied in pre-processing hydraulic properties for consistency.

The depths in each cell (d_i) are then computed using the closest DHAND information corresponding to the interpolated depth $\hat{d}_j$, and depths in each cell are computed as $d_i = \max(\hat{d}_j - H_i, 0)$. Blackbird supports the use of DHAND or HAND for the post-processing step of map generation, as well as the options to interpolate depths or apply the depth uniformly within the reach.

5.2.6 Evaluation metrics

Two key metrics are used to evaluate the quality of models against either a benchmark model or flood polygons. In the case where a benchmark model result with depths in each cell is available as the 'observed' data set, error metrics based on the difference in depths between cells may be computed. The first of these is the mean absolute error (MAE), computed as:

$$\text{MAE} = \frac{1}{N_m} \sum_{m=1}^{N_m} |d_m^s - d_m^o|$$

where N_m is the total number of cells evaluated (typically, all cells within a defined boundary), d_m^s for simulated depth in the mth cell and d_m^o as the 'observed' depth (noting that here we treat modelled depths from the 2D model as 'observed'). The range for this metric is $[0, \infty]$, where 0 is an ideal score.

The second metric used in evaluating errors in depths is the normalized sum of squared errors. This is computed as a sum of squared errors divided by the sum of errors when compared to a raster with the overall average depth mapped to each cell:

$$\text{NSSE} = 1 - \frac{\sum_{m=1}^{N_m} (d_m^s - d_m^o)^2}{\sum_{m=1}^{N_m} (\bar{d_o} - d_m^o)^2}$$

where $\bar{d_o}$ is the average depth in the evaluated cells within the observed raster ($\frac{1}{N_m} \sum d_m^o$). A score of one means the simulated result is equivalent to the naïve model of taking a single mean depth in all inundated cells.

Other relevant metrics related to the binary classification of images based on true positive, true negative, false positive, and false negative are often used in flood studies, such as hit rate, false alarms, error bias and Matthew's Correlation Coefficient (MCC) (e.g., *Wing et al.*, 2017; *Chicco*, 2017). These are omitted here in favour of the depth comparison metrics, since the available data supports additional information provided by specific depths.

5.2.7 Case studies

We present two case studies for verification and testing of the Blackbird approach. First, we provide the 'Simple Benchmark' case study, which is comprised of a set of six test cases (SB01-SB06), which are intended to verify the ability of Blackbird to provide a steady-state solution using the standard step method with cross-sections before it is used in the reach-integrated configurations. We deploy HEC-RAS 1D and Blackbird

to both model sets of cross-section data without any reach-integration or other novel methods presented in this manuscript.

The second case study is based on a physical location near the community of Waldemar, Ontario within the Grand River watershed. The base terrain is based on high resolution bathymetric and topographic data to ensure representation of the channel and floodplain geometries. In this case study we apply a fully 2D quasi-steady state model in HEC-RAS, which is treated as the 'truth' and used as the point of comparison for the 1D-based approaches. The purpose of this case study is to compare the HAND-Manning method, a cross-section based 1D model in HEC-RAS 1D, and the Blackbird model in their ability to approximate the results from the fully 2D model. The model results for both uncalibrated and calibrated models are provided.

The HAND-Manning method is also implemented within the Blackbird software, and uses the reach-integrated estimation of hydraulic properties from a HAND raster that is otherwise consistent with the Blackbird methodology. The normal depth in each streamnode is then estimated using Manning's equation to determine depth, providing the reach-integrated bed slope as the slope term. The depth results are generated for each streamnode and combined to create a flood map. The HEC-RAS 1D model refers to a fully 1D model built with HEC-RAS software offered by the US Army Corps of Engineers (*Brunner*, 2022). The same software is used to interpolate results between cross-sections and generate flood maps.

Simple Verification Benchmarks

In the Simple Verification Benchmark case study, we deploy six basic tests in comparing cross-section based models to ensure that the Blackbird standard step approach is consistent with industry-standard software. We apply the following tests in order of increasing complexity:

1. SB01 - basic trapezoidal channel with symmetric roughness values in the left/channel/right banks, consistent length values of 50m for left/channel/right
2. SB02 - basic trapezoidal channel with asymmetric roughness values (higher roughness in one bank than the other)
3. SB03 - basic trapezoidal channel, asymmetric roughness values, asymmetric and varying channel lengths in banks
4. SB04 - basic trapezoidal channel, varying channel lengths in banks, horizontally varying roughness values as a vector rather than specifying 3 roughness values for left/channel/right
5. SB05 - basic trapezoidal channels on three reaches, asymmetric roughness values, asymmetric and varying channel lengths in banks, and a junction.
6. SB06 - 49 realistic cross-sections cut from real terrain data along a single channel with horizontally-varying roughness values.

The last test, SB06, uses the sections and setup from the Waldemar 1D case study (described in the following section) under the high flow scenario. These six tests are used as a verification of the ability of Blackbird to correctly compute steady-state subcritical hydraulic results as a necessary check before deploying those algorithms for more complex application. Each set of results uses the depth at each cross-section to verify the agreement of the models.

Waldemar

This case study is developed as a test bed to compare the various modelling approaches against the results of a 2D hydraulic model (built using HEC-RAS 2D). The philosophical approach taken here is to consider the 2D model as 'observed' or 'truth' data, with the goal of the simplified modelling approaches being to approximate the 2D model as closely as possible.

The town of Waldemar is located along the Grand River in Ontario, approximately 14km west of Orangeville. The Grand River flows from the north through the town of Grand Valley, meanders and then flows through the town of Waldemar. The extents of the case study are depicted in **Figure 5.3**. The terrain was generated from a combination of the Southwestern Ontario Orthophotography Project 2020 (SWOOP2020) data product and bathymetry data provided by the Grand River Conservation Authority (GRCA). The SWOOP2020 product is freely available at a 0.5m resolution across all of southwestern Ontario, including the Waldemar study area, and is offered under the

Open Government License - Ontario. The bathymetry data was provided as a point cloud file under the GRCA Open Data License v2.0. The bathymetry data was processed into a 0.5m DEM using the `lidR` package in R (*Roussel et al.*, 2020; *Roussel and Auty*, 2024). These products were merged into a single terrain file at a 2m resolution to capture both bathymetry and terrain elevations. Land cover data was obtained from the GRCA Land Cover 2017 product (also available under GRCA Open Data License v2.0) and reclassified to Manningn's n roughness values using standard values from *Chow* (1959).

The flows provided to the model were derived from a flood frequency analysis on annual peak daily flow data obtained from Water Survey of Canada gauge 02GA014. The data was corrected by a drainage ratio to the location of historic gauge 02GA022 located within the study area, and then analyzed for extremes by fitting to the Generalized Extreme Values distribution using the `extRemes` package in R (*Gilleland and Katz*, 2016). The flows used in this case study were 156.2 m^3/s (low flow), 246.4 m^3/s (medium flow), and 314.7 m^3/s (high flow), which correspond to approximately a 1 in 2 year flow, a 1 in 10 year flow, and a 1 in 100 year flow (50%, 10%, and 1% annual exceedance probabilities), respectively. Only flows along the Grand River were considered in this model benchmark, though other tributaries exist within the study area.

The HAND-Manning, HEC-RAS 1D, HEC-RAS 2D and Blackbird models are compared within the Waldemar case study. All four models used the same terrain, land

cover, and flow inputs. Fifty cross-sections for the HEC-RAS 1D model were cut over the 13.5km domain and adjusted to avoid overlapping in meanders. The cross-section thalweg locations were used as streamnode locations for the HAND-Manning and Blackbird models. Terrain and roughness data for the HEC-RAS 1D model was obtained from the raster inputs using native HEC-RAS tools to populate cross-section data. Bridges are present at crossings in the area but are omitted in all models here for simplicity. The roughness values in all non-2D models were initially multiplied by a single roughness multiplier of 1.1 in recognition that the losses encapsulated by the roughness term vary based on the modelling approach (*Morvan et al.*, 2008), and are generally higher in more simplified models where more energy losses are abstracted by the roughness coefficient rather than being defined in the mathematical formulation. The initial 'uncalibrated' factor of 1.1 was determined by comparing depths in key sections of the HEC-RAS 1D and 2D models, and adjusting the global roughness multiplier until the error in depths at those sections was minimized. The roughness multiplier was later calibrated manually for each of the three benchmarked models using the NSSE scores to develop calibrated models.

In developing the 2D hydraulic model with HEC-RAS 2D, a base cell resolution of 5m was used with additional refinement along the main channel and hydraulic features in the model, such as berms and ponds. The model was run in unsteady mode using the Shallow Water Equations Eulerian-Lagrangian Method (SWE-ELM) method. The Conservative turbulence model was also included to account for turbulent effects. The

timestep was configured as an adaptive timestep, with an initial timestep of 0.5s and allowing a range of 0.016s to 4s timesteps. The maximum depth error (as reported by the HEC-RAS solver) across all three flow simulations was found to be less than 0.23m in water surface elevation, with an overall volume accounting error of 0.0047% or less in each run. The model run time was between 2-4.5 hours per run on an i9 desktop with 24 cores, depending on the flow scenario. The HAND-Manning and Blackbird models required a pre-processing runtime of approximately 70min to compute hydraulic properties prior to the model runs. The model runtime of a few seconds was negligible in comparison, and is similar to the runtime in HEC-RAS 1D. The generation of result rasters using any approach ranged from 30-120 seconds.

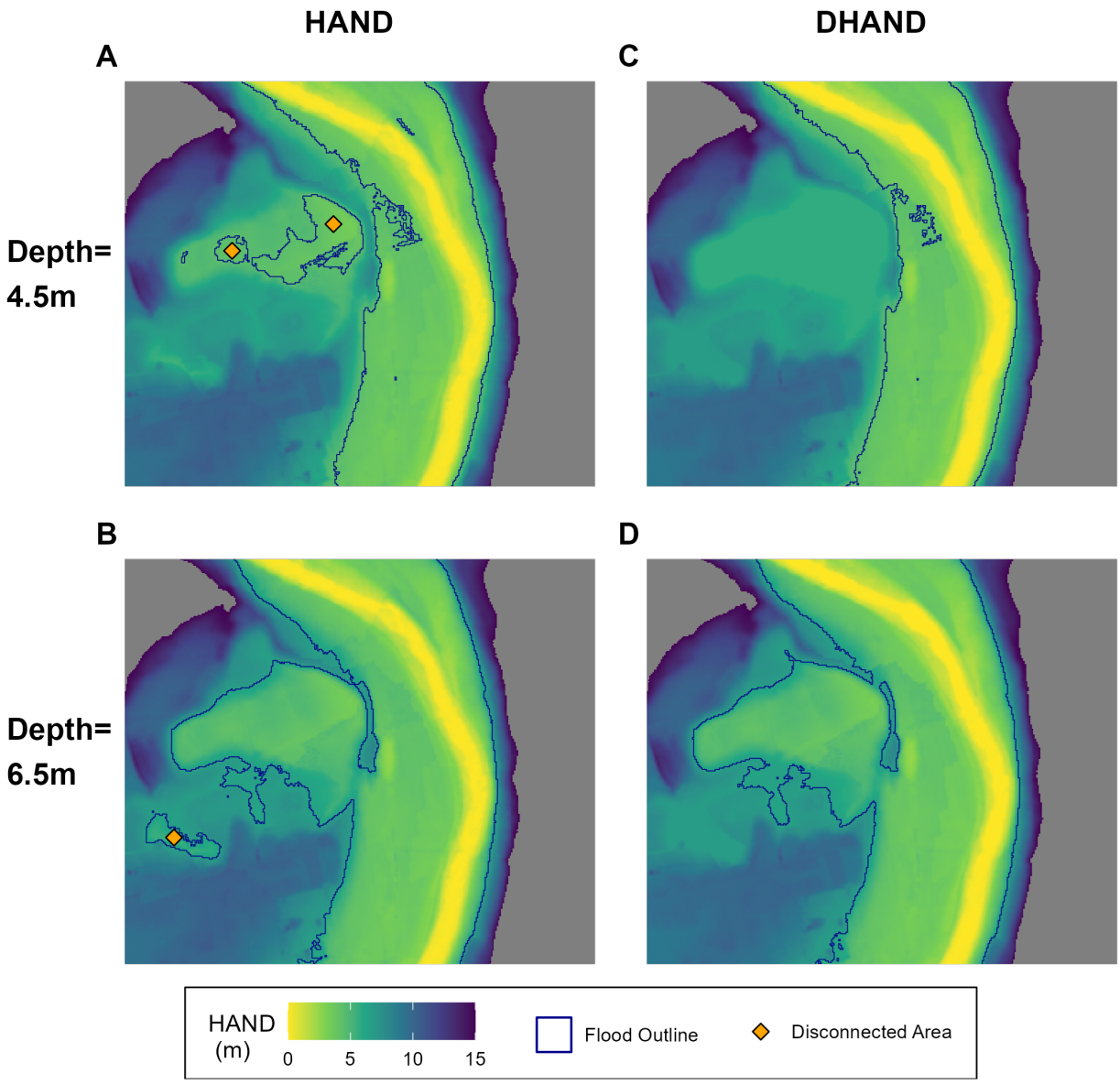


Figure 5.2 HAND and DHAND layers are applied for fixed depths to compute flood extents adjacent to a river bend. A) HAND layer with shallower depth of 4.5m shows flooding in two hydraulically disconnected areas on the inside of a berm. B) Same HAND layer from plot A with a greater depth of 6.5m shows greater flooding inside the berm area as it is overtopped, and also shows a new disconnected and inundated area appears just southwest of the original area. C) A DHAND layer for the specific depth of 4.5m shows no flooding on the other side of the high berm. D) The DHAND layer for a depth of 6.5m now allows flooding on the other side of the berm as it is overtopped, and does not generate falsely inundated areas as with the HAND layer in plot B.

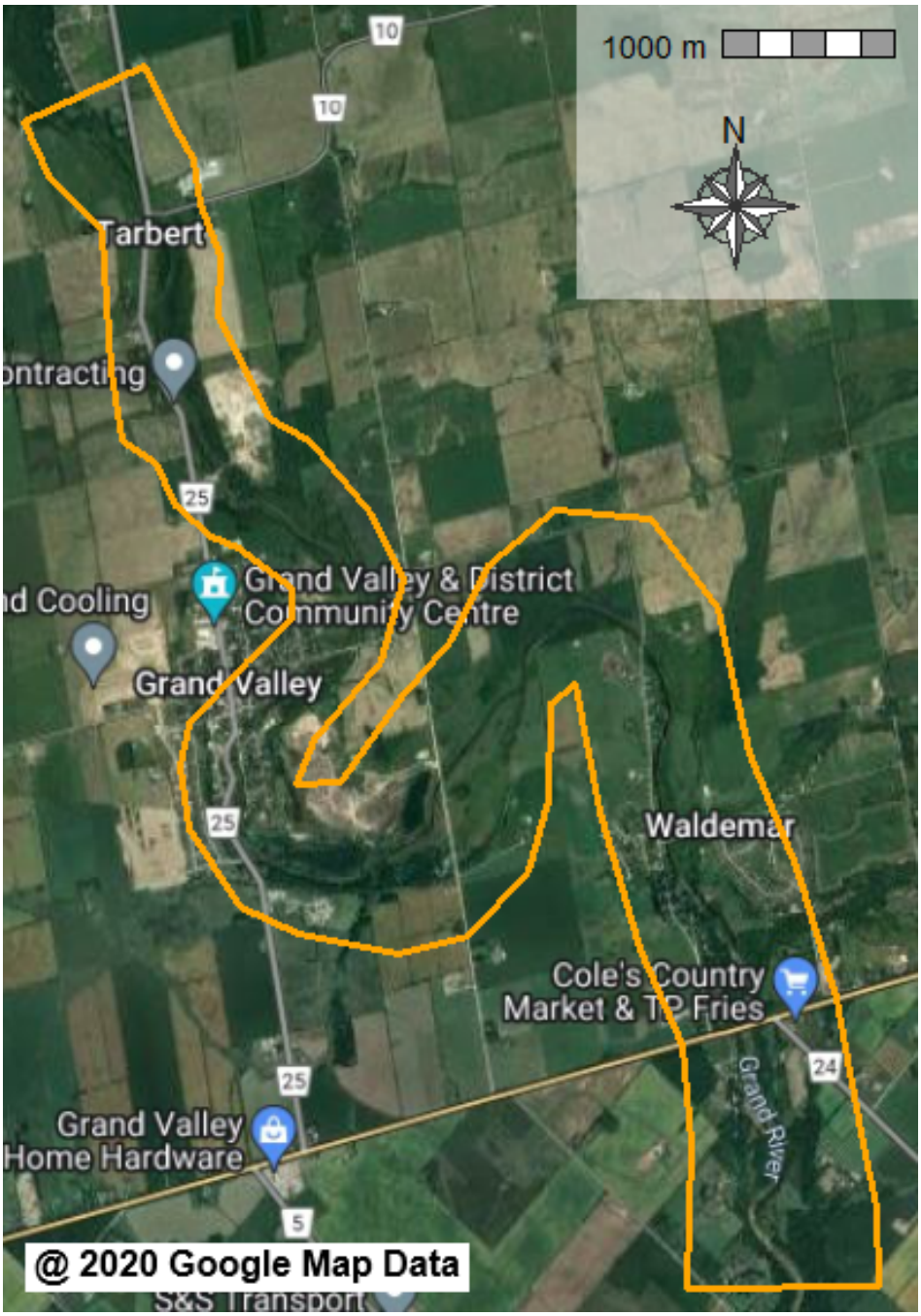


Figure 5.3 Location of the case study in the Town of Waldemar. The extents of the HEC-RAS 2D model are shown in the orange outline. Imagery obtained from Google 2020 imagery.

5.3 Results

5.3.1 Simple Verification Benchmarks

The results from the Simple Benchmark tests found that in case studies SB01-SB05, the Blackbird algorithm had a maximum error in depth of 1cm (translating to approximately 0.08% error in depth), which was found on a few sections and is likely due to small numerical errors or rounding to the nearest centimetre. Aside from that, the HEC-RAS 1D results are effectively emulated by the Blackbird code for these simple tests.

In SB06, the full complexity of model inputs derived from real data including up to 500 ordinates per cross-section, rather than the theoretical cross-sections and other parameters (such as reach lengths and roughnesses) used in SB01-SB05. The maximum error in SB06 was bounded by ± 0.2m, which is approximately $\pm 5.25\%$. Based on the results of a t-test on the mean and Shapiro-Wilk test respectively, the mean of the data is not significantly different from zero and appears to be normally distributed. A plot of the errors is provided in **Figure 5.4**, indicating that the errors appear randomly distributed along the channel station (i.e. not accumulating with station) and normally distributed. The errors in SB06 may be due to aggregated rounding errors in each cross-section, as the modelled cross-sections have between 175 and 500 points each, while earlier experiments had only eight (8) points per cross-section. There may be other

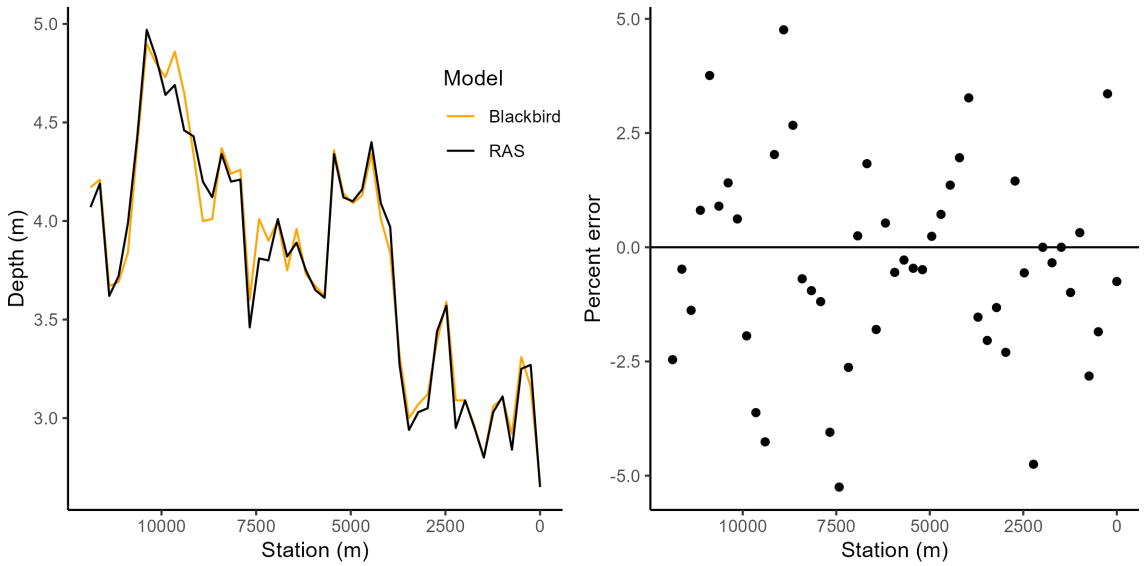


Figure 5.4 Results of the SB06 benchmark test, summarized with a profile plot for both Blackbird and HEC-RAS generated results on the left and the residual depth errors shown on the right. Errors are bounded by ±0.2m and are randomly distributed around a mean of zero.

differences in the algorithms as well, such as in the algorithmic handling of inundated areas outside of the main channel or other nuances that were not represented in SB01-SB05.

Overall, the simple benchmark results indicate that the Blackbird code is working as expected when deploying the standard step method in subcritical regimes for a set of cross-sections when compared to HEC-RAS 1D. The code is able to replicate the results of simple case studies almost exactly, and with a small amount of error in a very complex terrain. This provides sufficient confidence for using the code in the reach-integrated application, which is the focus of this study.

5.3.2 Waldemar

The MAE and NSSE metrics (defined in **Section 5.2.6**) were computed for each scenario and flow in the Waldemar case study are provided in **Figure 5.5**. The evaluation is based on all raster cells within a defined evaluation area, and for each flow regime (low, medium, or high), any cells that are consistently dry in all model results are removed from the calculation of metrics. This pruning of dry cells does not impact the relative ranking of models in the results. The results indicate that based on all metrics, the best model in each flow scenario was consistently Blackbird, followed by HEC-RAS 1D and then HAND-Manning.

The HAND-Manning method is consistently ranked last in performance with errors substantially larger than other methods. The depths generated by HAND-Manning methods are always underestimated in this case study, and require calibration to increase roughness by a large factor in order to approach reasonable results. This in particular highlights the value added by considering energy losses via a hydraulic model, rather than performing geospatial-based analysis alone, as the energy loss impacts are substantial.

Results were all improved by calibration, though the HAND-Manning method was substantially improved and the Blackbird results were only marginally improved as the initial roughness multiplier was almost unchanged. The calibrated roughness multiplication factors were 4.83, 1.35, and 1.14 for the HAND-Manning, HEC-RAS1D and

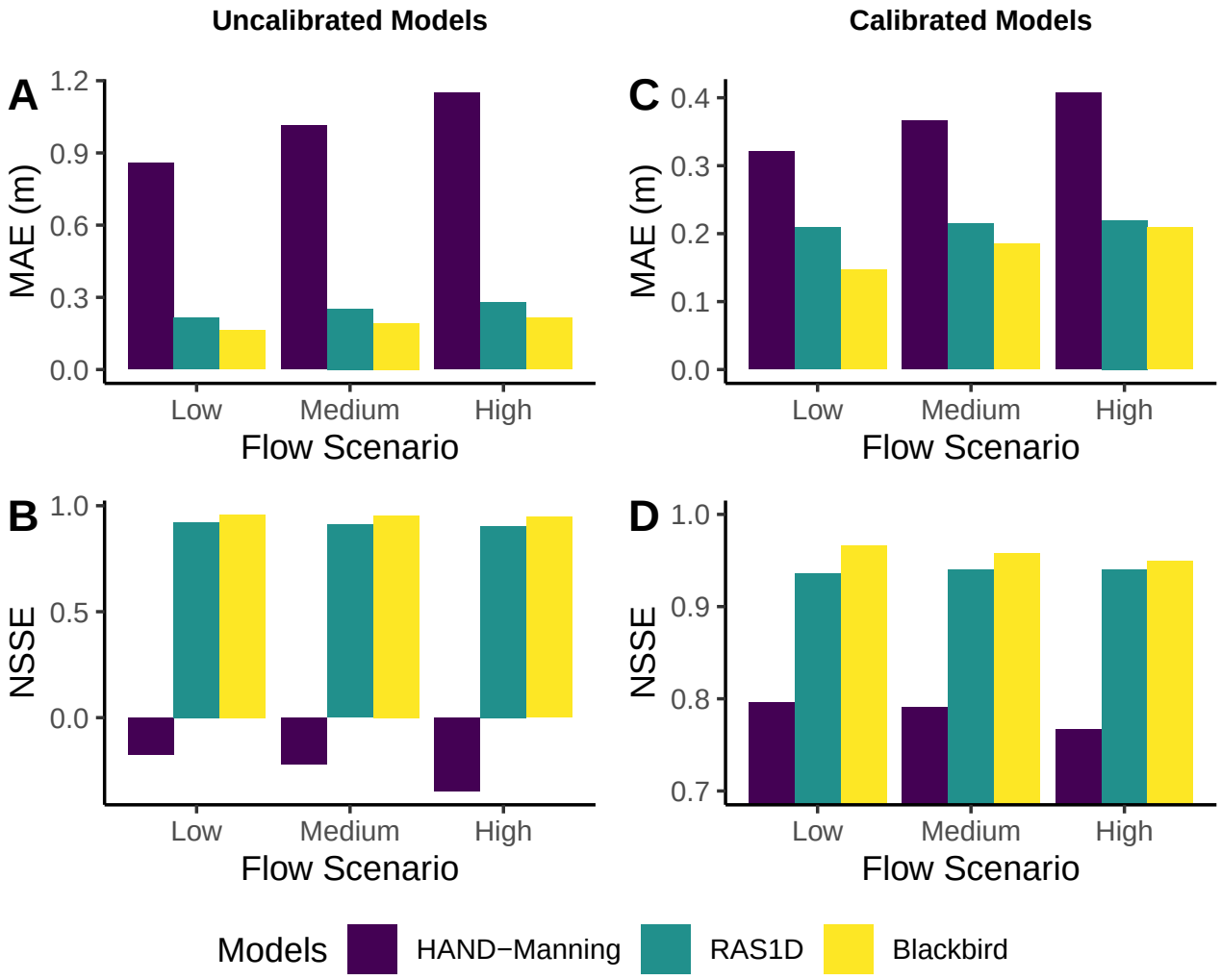


Figure 5.5 Results for the three model runs evaluated against the HEC-RAS 2D model under three flow scenarios. All raster cells within a defined evaluation area were evaluated, and cells that were not consistently dry in all four models (including the 2D model) were used in the calculation of metrics. Metrics used are the mean absolute error (MAE) and normalized sum of squared errors (NSSE). Plots A and B show uncalibrated model results, and plots C and D show calibrated model results.

Blackbird methods, respectively (compared to an initial multiplier of 1.1). The difference in roughness multipliers highlights the drastic need to compensate for a lack of inherent energy losses in the HAND-Manning method. The difference in calibrated multipliers between the HEC-RAS 1D and Blackbird models also suggests that there may be additional energy losses captured with the reach-integrated approach relative to the cross-section approach, leading to the lower multiplier in the Blackbird model. Different model formulations leading to different calibrated Manning's roughness values (*Morvan et al.*, 2008).

The model results were visually compared to the 2D model results in **Figure 5.6** to show locations where the inundated (wetted) and dry areas varied from the 2D output. Outputs were mapped in a categorical classification scheme to show where results were 1) accurately classified as wetted (Hits), 2) false positives, where the model produced a wetted cell in a cell with no water from the 2D model, and 3) false negatives, where a dry well was produced where the 2D model was wetted.

Figure 5.6 shows clearly how the HAND-Manning method has the largest swaths of both false positive and false negative areas. The area on the inside bend near the northern part of the study area is wetted (false positive) in both the HAND-Manning and HEc-RAS 1D results though not the Blackbird results, a phenomenon which is further explained by **Figure 5.2**. While some areas of both false positives and false negatives exist in all three model results, the Blackbird model is successful in minimizing the large areas of false positives due to channel connection issues, or

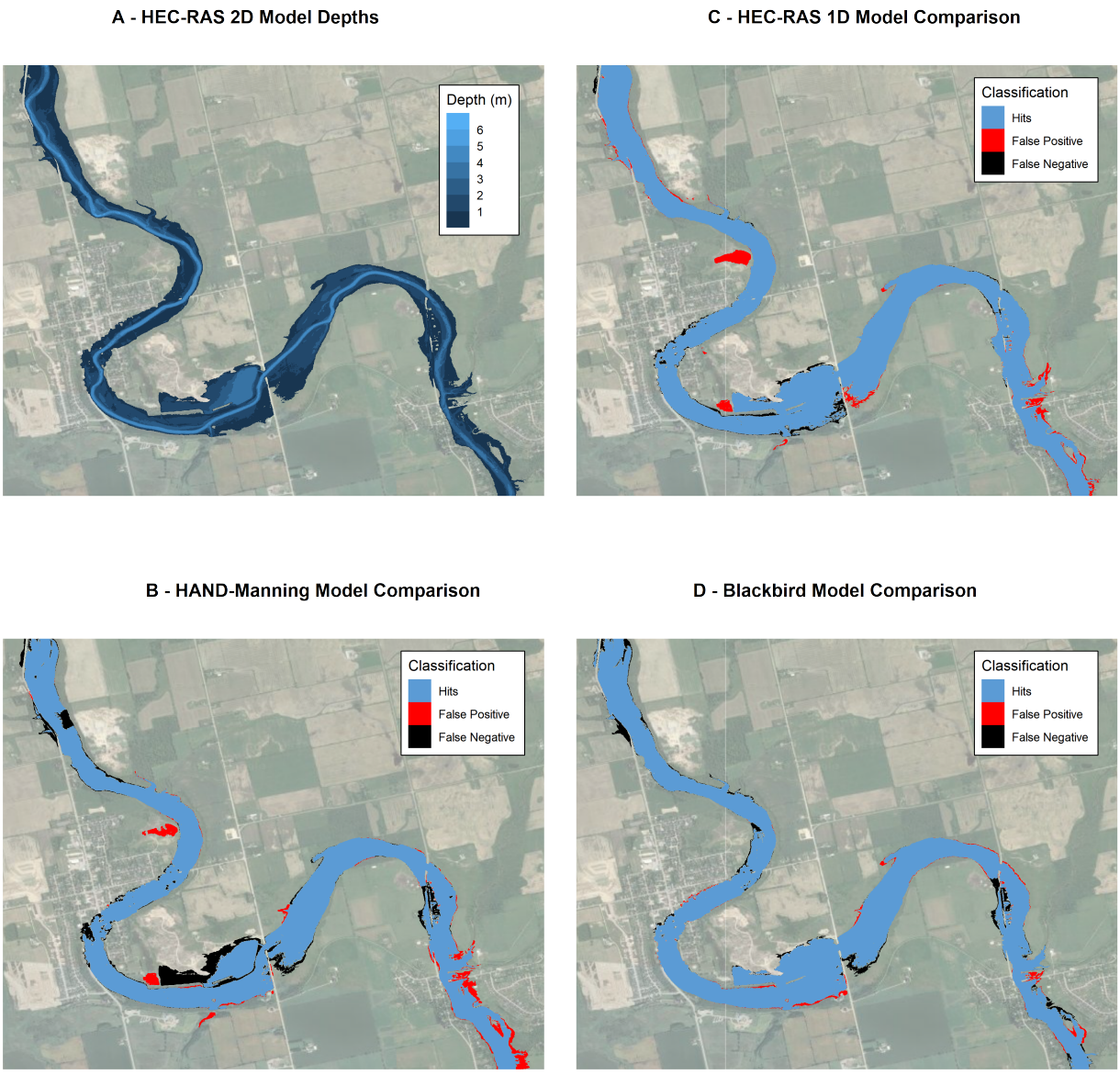


Figure 5.6 Calibrated model results compared against 2D model results, showing the results for (A) the HEC-RAS 2D model depths, (B) the HAND-Manning model classified differences, (C) classified differences for the HEC-RAS 1D model, and (D) classified differences for the Blackbird model output.

from resolving the threshold at which the channel spills to a low lying area. This can be attributed to a combination of accurate depth estimations as well as the DHAND method used to resolve spatial connections on the landscape.

5.4 Discussion

5.4.1 Inundation forecasting requirements

Computational runtime is an important consideration both in large-scale and forecasting applications, where obtaining accurate model results in a timely fashion is critical. This may be compounded if any model calibration or uncertainty analysis requires the model to be run hundreds or thousands of times in a short timeframe, and potentially without the computing resources necessary to run 2D models quickly.

A key benefit of Blackbird is that the computationally expensive estimation of hydraulic properties is handled in pre-processing, leaving the remaining simulation runtimes to be essentially equivalent to that of any 1D model. For instance, in the Waldemar case study the simulation runtime of the HEC-RAS 2D model runtime was 4.5 hours for the high flow scenario while the Blackbird runtime was less than a few seconds, which is approximately $\times 10,000$ faster. This difference in model run time can be expected to increase substantially with studies of a larger extent. It is also worth noting that the same pre-processing is required for the HAND-Manning method as Blackbird, thus despite the simplicity of the HAND-Manning technique the

computational cost savings are negligible.

A feasible current practice for estimating flood extents in real-time forecasting applications is through the use of existing flood maps. Agencies responsible for flood forecasting will typically possess regulatory maps of estimated flood extents for events with a specific probability, defined by an annual return period or exceedance probability. For example, a forecaster may access maps showing the flood extents of the 1 in 50 year and 1 in 100 year flood events for an area of interest. A real-time flow estimate may coincide with a 1 in 80 year return period, in which case the two maps for the 50 and 100 year extents would be used to estimate flood extents of the current event. However, complications arise in this kind of interpretation that can render this exercise very challenging. For example, the maps may be out of date in either the estimated probabilities of flows, the landscape information, or the methodologies which produced the maps, leading to complications in easily interpreting flood maps under the pressure of issuing forecasts and decisions and real-time. At junctions, the regulatory flood maps are often made with idealized scenarios regarding the timing of peak flows or equal probability of flows at both upstream sections of the junction (e.g., 100 year flow on each branch). However, in practice there will likely be a different probability determined for the flows on each branch, such as a 1 in 80 year flow on one branch and 1 in 130 year flow on the second branch, making the use of a set of regulatory flood maps less useful. There may also be backwater effects impacting flooding due to major constrictions on the channel downstream of the area of interest, which may not have

been considered in the generation of the regulatory map. In these scenarios the regulatory flood maps are far less useful to forecasters and decision-makers than a model which can be run in real-time. Inundation forecasting, i.e. issuing forecasts of flood depths at specific locations, is much less common today than streamflow forecasting, but may be supported by the capabilities of Blackbird to address many of the above issues and support accurate inundation maps in real-time.

5.4.2 Relative model development effort

Another important consideration in developing models is the relative effort requiring manual intervention. Construction of cross-section based 1D models is often labour intensive, requiring manual specification of bank stations at each cross-section, defining left/right/channel lengths, aligning cross-sections to avoid overlapping, adjustment of sections to ensure a smooth interpolation between sections, and redefining these values (including lengths, elevations and roughness values) with each re-alignment. In addition, special attention must be given to the location of cross-sections relative to the channel profile, and ensuring that sections are consistently placed at the top of riffles or other consistent features within the profile. Inconsistent placement of sections along the profile can lead to a poor definition of the channel inverts solely as a result of cross-section placement. Two-dimensional models also involve a fair amount of manual work, particularly in defining the breaklines during mesh generation, which are required to avoid artificial 'spilling' of water across cells.

By contrast, Blackbird simplifies much of the effort to a set of automated functions, and the only manual labour is the same quality assurance required for any modelling exercise. For example, cross-sections do not need to be cut; only point locations for computational nodes need to be defined, which can be done automatically and are theoretically less sensitive to placement along the channel profile than cross-sections because of the reach-integration approach. Contributing area extents are defined by catchment areas around each streamnode. These still require checking but are generated automatically and therefore less prone to error and bias by a modeller, such as extending a cross-section short of the potential flood extent. Bank stations and reach lengths are not explicitly defined in the Blackbird configuration as they are in HEC-RAS 1D, as bank stations are not used and reach lengths are calculated dynamically. Overall, the degree of manual effort is greatly reduced in the development of a Blackbird model relative to the traditional cross-section based 1D model or fully 2D models.

5.4.3 Modelling challenges and limitations

The popularity of HAND-based flood maps has increased in recent years, which is likely driven by the ease of implementing the method as well as the low computational requirements. Validation of HAND-based models is often presented as adequate when the HAND-generated flood map captures some of the flood extents derived from satellite imagery (e.g., *Scriven et al.*, 2021; *Chaudhuri et al.*, 2021), or in comparison to

other even simpler methods (e.g., *McGrath et al.*, 2018). However, the limitations of this method are sometimes downplayed and are worth further discussion. The HAND-Manning method applies Manning's equation for depth estimation in each defined reach independently, and thus depths may vary substantially between reaches without any impact on adjacent reaches. The HAND-Manning is noted to underperform in certain environments, such as meandering portions of channels or near hydraulic structures (*Afshari et al.*, 2018), and is considered by *Jafarzadegan et al.* (2023) to only work in confined floodplains. The inability of the HAND-Manning method to incorporate energy losses, hydraulic structures, ice jams, or backwater effects is also severely limiting, as shown partially by the results of this study. While idealized conditions may be encountered, often the events that cause severe flooding are related to one or more of these hydraulic conditions (i.e., high energy losses, backwater effects from a downstream constriction, obstructions at bridges or culverts, ice jams, etc.). A model that is validated under idealized flows and then applied to forecast or simulate a real flood event under one of the aforementioned conditions may result in severe underprediction of depths with potentially disastrous consequences.

In comparing Blackbird with HEC-RAS 1D, the models are fundamentally similar in the computation because of the exact same underlying sets of equations and solvers (under steady state conditions). Where the models differ is in the landscape discretization and representation within the model. Theoretically, the GASS method uses more of the available information in determining hydraulic properties than a cross-section

approach, which discards the majority of data available to the modeller. It is unlikely that there would be scenarios in which GASS would underperform the cross-section methodology under the same set of input data. The scenarios where this may not be true are primarily related to the underlying catchment delineation algorithm used to define reach areas, such as very flat areas where catchment delineation is a challenge, or other scenarios where the reach-integration approach is not a fair reflection of the hydraulic conditions. However, the ability of Blackbird to also represent streamnode properties with cross-sections alleviates any limitations relative to the cross-section approach, and future developments planned for Blackbird (such as representing hydraulic structures) will likely remove this relative limitation altogether. Other challenges of 1D modelling, such as conditionally varying or ephemeral flow paths (i.e. flow paths that may 'short-circuit' the channel for some depths) are a challenge for both Blackbird and HEC-RAS 1D, since both of these models require a defined static channel line. Such cases are currently best handled with 2D models.

The Blackbird and HEC-RAS 1D models presented here are both steady-state models, which assume that flow values are sustained for a sufficient amount of time for steady-state conditions to be achieved. In practice, this is a conservative assumption compared to unsteady state where a time series of flows is provided rather than a fixed peak value. Under unsteady flow, the volume of water provided by the flow boundary conditions is taken into account in the hydraulic modelling. This can be particularly important in low-slope or flat areas where a small change in average depth of a cross-

section of streamnode can result in vastly greater volumes of water on the landscape. As a check in the Waldemar case study, the cross-sectional flow area in each reach-integrated streamnode can be multiplied by the effective reach length, and summed to get the total volume of water in the modelled landscape. Dividing this by the flow and converting units reveals a time period of approximately 3.5 hours, which is conceptually the duration that the given flow rate would need to be maintained in order to deliver enough water to the system to fill the landscape (including floodplains) to the modelled extents. In the Waldemar example, a large proportion of this storage comes from a single streamnode where an offline storage pond is found and assumed initially dry in this model. This type of check on the system is important particularly when the landscape is flat, which can lead to overly conservative estimates of flood extents, and this duration check can reveal whether the steady-state assumption over the entire domain is a reasonable one.

5.4.4 Future development

Future developments are planned to further test and expand the capabilities of the Blackbird software. The further testing would involve fully understanding under which conditions the GASS approach performs well and where it does not, and what modifications may be required to address those limitations. Additional testing for features that are implemented though not studied in this work include the modelling of junctions, and including a mixed representation of streamnodes as cross-sections and reach-integrated

properties within a modelling domain.

Additional feature that are not yet implemented include the ability to represent hydraulic structures, such as bridges and culverts, which are very commonly encountered in systems with potential to impact humans under flood conditions. The development of hydraulic structures will take two approaches: 1) implementing hydraulic structures with cross-section based streamnodes akin to what is done in HEC-RAS 1D, and 2) in more abstract reductions in flow area at streamnodes, which would be less accurate but with fewer data requirements. The latter in particular is intended to support modelling at large scales, where the presence of hydraulic structures can be detected at road crossings through geospatial analysis, but detailed survey information may not be available.

Another planned extension is the inclusion of unsteady flow (i.e. 1D Saint Venant equations) to support time-varying hydrographs as boundary conditions. As discussed in **Section 5.4.3**, the steady-state assumption may be overly conservative, especially in flat areas, and lead to an overestimation of flood risk. This can be addressed by the use of unsteady flow solutions. The inclusion of ice jams is also a possible extension to Blackbird capabilities. As with hydraulic structures, ice jams could be represented with either a cross-section approach as is done in HEC-RAS 1D, or with a more generalized reach-integrated approach that reduces the flow area and increases resistance to flow in particular streamnodes. The latter may be more promising as the locations and properties of ice jams are often a source of great uncertainty, and thus the ability to

reduce sensitivity to the precise location of ice jams may have great value in simulating ice jams under flood conditions.

5.5 Conclusions

A new method, GASS, has augmented the standard step method to handle complex terrain features and overcome interpolation issues encountered while generating flood maps using traditional 1D hydraulic models. The GASS method, as implemented in the Blackbird software, maintains computational efficiency during runtime by preprocessing the required hydraulic variables, and also substantially reduces the effort required to develop hydraulic models for a given location of interest.

Blackbird was first successfully verified in its ability to emulate the steady-state solver in HEC-RAS 1D with cross-section models, and was then compared to existing modelling approaches in a benchmark case study near the Town of Waldemar, ON. The benchmark study shows that Blackbird vastly outperforms the popular HAND-Manning approach and can also improve the accuracy of inundation dpeths and locations relative to existing 1D models, partially due to its improved handling of landscape data through DHAND. The Blackbird model with its DHAND approach was shown to be particularly adept at minimizing the occurrence of false positives resulting from inadequate handling of landscape connections in comparable HAND or 1D models. The importance of embedding energy loss calculations in the models was demonstrated by the severely degraded performance and underestimation of depth by the HAND-

Manning approach compared to either 1D hydraulic modelling approach evaluated in this study. This also highlights the significance of representing hydraulic structures and other 1D features through models such as Blackbird, which is not possible with the HAND-Manning method.

Blackbird has the potential to advance the state of flood mapping and inundation forecasting, particularly at larger spatial scales. The computational speed of Blackbird was on the order of $10,000$ times less than the 2D hydraulic model in the Waldemar case study. This illustrates the feasibility of deploying this methodology at spatial scales in which 2D hydraulic models are too computationally expensive to deploy and when model results are needed in a timely manner, such as forecasting. Future improvements to Blackbird are likely to include features such as unsteady flow and hydraulic structures. The ability of Blackbird to support a mixed representation of streamnodes as either cross-sections or reaches means that these developments may utilize existing approaches and equations. Overall, the GASS method as implemented in Blackbird software includes all of the benefits of 1D models, such as fast runtimes, while addressing some of the shortcomings of existing flood mapping techniques, including the high degree of model development effort and the imputation of depths to the landscape without preserving appropriate hydraulic connectivity to the channel. With these innovations, Blackbird provides an improved option to support real-time and large-scale flood mapping applications in the future.

Chapter 6

Conclusions

6.1 Conclusions

This thesis addresses three main objectives. Here, we present how each thesis chapter corresponds to these objectives and key findings.

The first thesis objective is **to test the use of the blended hydrologic model in the context of model calibration, and demonstrate its potential to improve both overall model performance and the identifiability of model structure**. Chapter 3 addresses this objective, and demonstrates both that the overall model performance and efficiency of scientific inquiries, such as hypotheses on dominant model structures, can be improved by use of the blended model structure. The computational benefits are also demonstrated, as the blended model uses a single calibrated model structure relative to the alternative approach of generating many ensemble members

(108 in this study) and calibrating each model structure individually. The performance of the blended model is shown to be greater than any of the 108 plausible structures as well as the ensemble models, which require the generation of each ensemble members in order to average the results in some form.

The second objective is **to develop, apply, and assess an updated blended model configuration with improved performance when applied across catchments within the contiguous USA**. Chapter 4 addresses this objective by experimenting with the original blended model structure, and iteratively developing a new blended model structure with superior performance across the contiguous USA. For example, the new blended model structure achieves a mean KGE score across the 12 MOPEX catchments of 0.9 in calibration and 0.8 in validation, which represents a relative improvement over the original model from a KGE score of 0.84 in calibration and 0.73 in validation. This updated model structure, referred to as the blended v2 model, can be deployed 'out-of-the-box' in a variety of hydrologic regimes and achieve high performance in calibration, which was the goal of the study.

The third and final objective is **to develop and assess a new method for efficient physically-based floodplain mapping that may be useful in large-scale and real-time applications**. Chapter 5 addresses this objective by presenting the development of the Geospatially Augmented Standard Step (GASS) method, a novel approach to hydraulic modelling and mapping which merges a 1D hydraulic model with an improved version of the geospatial HAND method. The GASS method is im-

plemented in a software package called Blackbird and is first verified to successfully replicate the standard step method against HEC-RAS 1D, which is shown to exactly emulate results for simple tests and achieve a hydraulic profile within $\pm 0.2m$ of HEC-RAS 1D results for a complex set of cross-sections. Blackbird is then benchmarked for performance in floodplain mapping against the HAND and HEC-RAS 1D models, using a HEC-RAS 2D model as the 'truth' for evaluation purposes. The experiment shows that the Blackbird model vastly outperforms the HAND model and also outperforms HEC-RAS 1D, and is particularly useful in removing the occurrence of false positives in the inundation mapping. The workflow for developing Blackbird models also removes much of the manual effort normally required configure 1D models while maintaining a simulation runtime on the order of $10,000$ less than that of 2D models in the benchmarked domain.

In summary, this thesis contributes to the development of tools for the efficient development and deployment of hydrologic and hydraulic models in support of water management activities, including large-scale mapping and real-time forecasting of both streamflow and inundation. The hope is that these tools will benefit practitioners in producing high performing hydrologic models and generate flood maps in locations where they have historically been infeasible or cost prohibitive to generate.

6.1.1 Limitations

The work presented in this thesis is subject to some limitations. The application of the blended model presented in the thesis is applied only in lumped configurations. Minor modifications to the model structure as well as further testing would be required to apply the model in semi-distributed mode, which would include some decisions about which parameters to divide within land use, vegetation or soil classes. A semi-distributed version of the original blended model was applied in *Mai et al.* (2022a), though a careful application of the blended v2 model in semi-distributed mode is not yet available in the literature.

The results of Chapter 4 showed that the number of model parameters did not degrade model performance in calibration within the particular experiment, though a large expansion of the number of parameters in semi-distributed mode may challenge that. Evidence of this is partially provided by the conglomerate model experiment, in which substantially more parameters and options are added to the model and performance degrades. Thus, a limit does exist in the number of parameters and the amount of complexity that can be added to the model before performance degrades, which is partially a function of the calibration budget. In addition, both Chapter 3 and Chapter 4 are primarily tested and applied within the contiguous United States with a focus on the MOPEX 12 catchments. Additional testing for applications outside of these areas is highly recommended, since processes such as orographic corrections and cold regions processes which may not have been well captured in these lumped

model configurations may become quite important in other watersheds. Finally, the hydrologic models are heavily influenced by the input data, and the configuration in both blended model experiments relies only on temperature and precipitation data as inputs. With additional reliable input data types provided to the blended model, such as gridded solar radiation, wind speed, and humidity, it may be possible that more sophisticated process algorithms that rely on accurate estimates of these parameters may become more dominant in the application of the blended hydrologic model, even if they did not perform well in the experiments presented here. Changes to the input data, as well as other changes such as the metric for evaluating performance, may be reasons to consider checking the performance of other process algorithms beyond the blended v2 structure in future deployments of the blended model.

The hydraulic modelling and mapping chapter (Chapter 5) is also subject to limitations. As discussed in the chapter, the method is a relatively simple standard step method which assumes subcritical mode in solving hydraulic equations and also assumes steady state, which may underperform under conditions such as more complex hydraulic conditions involving hydraulic jumps, weirs, supercritical flow, and relatively flat areas of floodplain in which the steady state assumption is overly conservative. While it is not anticipated that the GASS method would underperform relative to a comparable 1D cross-section based model due to its ability to emulate the cross-section approach, the key limitation of the GASS method is the definition of catchment boundaries for each reach, which are subject to proper processing of terrain data and may

produce poor results under some drainage conditions. As with any new method, there remains much more testing to understand under which conditions the method works well, and in particular, under which conditions the method works poolry and requires modification to improve performance.

6.1.2 Future Work

Future work for the blended model should involve deployment of the blended model version 2 structure in semi-distributed mode, as well as to regions outside of the contiguous USA. This would likely involve testing other process algorithms, such as cold regions or orographic processes. Changes in input data to the setup, such as gridded meteorological inputs, would also trigger some re-examination of the processes included within each process group of the blended model. Other processes may also be blended in future experimentation, as the blending is not limited to the process groups used thus far. Experimentation with various objective functions that focus on high flow, low flows, or seasonal regimes would also enhance the use of blended models for scientific inquiries and process identifiability, as the relative identifiability of process options within process groups (e.g., highly identifiable infiltration and much less identifiable baseflow) is almost certainly a function of the objective function used in calibration.

Future development of the Blackbird software will include modifications to represent hydraulic structures such as bridges and culverts, which was not done in the benchmarking to date but is commonly encountered in applications of hydraulic mod-

els. The model may also be tested under conditions of ice formation and flooding due to ice jams, which is also commonly encountered in parts of Canada and has been implemented within other 1D hydraulic modelling softwares, and could therefore also be added within the Blackbird model. Features that are currently implemented in Blackbird but not fully tested include modelling junctions and use of a mixed representation of streamnodes via cross-sections and reach-integrated properties.

Future work to deploy Blackbird to larger scales will also test the design of the workflow in its stated ability to scale efficiently. While the software was designed with this feature in mind, the current benchmark experiment is still on the order of a local scale in which 2D hydraulic models can be feasibly run. Deploying Blackbird on much larger scales is the intended purpose of the model, but will likely reveal some additional challenges not encountered in smaller scales (such as data availability and efficient encoding of features in the model). The same may be also said of forecasting and real-time applications, where future work will likely focus on the optimization of the Blackbird software to maintain high computational speeds at scale for those specific application needs.

References

Addor, N., and L. A. Melsen (2019), Legacy, rather than adequacy, drives the selection of hydrological models, *Water resources research*, *55*(1), 378–390.

Addor, N., A. J. Newman, N. Mizukami, and M. P. Clark (2017), The camels data set: catchment attributes and meteorology for large-sample studies, *Hydrology and earth system sciences*, *21*(10), 5293–5313.

Afshari, S., A. A. Tavakoly, M. A. Rajib, X. Zheng, M. L. Follum, E. Omranian, and B. M. Fekete (2018), Comparison of new generation low-complexity flood inundation mapping tools with a hydrodynamic model, *Journal of hydrology (Amsterdam)*, *556*, 539–556.

Andréassian, V., A. Valéry, C. Perrin, B. L., and L. Moine (2009), Crash tests for a standardized evaluation of hydrological models, *Hydrology and Earth System Sciences*, *13*(10), 1757–1764.

Apel, H., G. T. Aronica, H. Kreibich, and A. H. Thieken (2009), Flood risk analyses–how detailed do we need to be?, *Natural hazards (Dordrecht)*, *49*(1), 79–98.

Aristizabal, F., F. Salas, G. Petrochenkov, T. Grout, B. Avant, B. Bates, R. Spies, N. Chadwick, Z. Wills, and J. Judge (2023), Extending height above nearest drainage to model multiple fluvial sources in flood inundation mapping applications for the u.s. national water model, *Water resources research*, *59*(5).

Arsenault, R., A. Poulin, P. Côté, and F. Brissette (2014), Comparison of stochastic optimization algorithms in hydrological model calibration, *Journal of hydrologic engineering*, *19*(7), 1374–1384.

Arsenault, R., F. Brissette, J.-L. Martel, M. Troin, G. Lévesque, J. Davidson-Chaput, M. C. Gonzalez, A. Ameli, and A. Poulin (2020), A comprehensive, multisource database for hydrometeorological modeling of 14,425 north american watersheds, *Scientific Data*, *7*(243), 2052–4463, doi:10.1038/s41597-020-00583-2.

Arsenault, R., J.-L. Martel, F. Brunet, F. Brissette, and J. Mai (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, doi:10.5194/hess-27-139-2023.

Asadzadeh, M., and B. Tolson (2009), A new multi-objective algorithm, pareto archived dds, in *Proceedings of the 11th Annual Conference Companion on genetic and evolutionary computation conference*, GECCO '09, pp. 1963–1966, ACM.

Asadzadeh, M., and B. Tolson (2013), Pareto archived dynamically dimensioned search with hypervolume-based selection for multi-objective optimization, *Engineering optimization*, *45*(12), 1489–1509.

Asadzadeh Esfahani, M. (2012), Developing parsimonious and efficient algorithms for water resources optimization problems.

Bates, P., and A. De Roo (2000), A simple raster-based model for flood inundation simulation, *Journal of Hydrology*, *236*(1), 54–77, doi:https://doi.org/10.1016/S0022-1694(00)00278-X.

Bates, P. D. (2022), Flood inundation prediction, *Annual Review of Fluid Mechanics*, *54*(Volume 54, 2022), 287–315, doi:https://doi.org/10.1146/annurev-fluid-030121-113138.

Becker, R., A. Koppa, S. Schulz, M. Usman, T. aus der Beek, and C. Schüth (2019), Spatially distributed model calibration of a highly managed hydrological system using remote sensing-derived et data, *Journal of hydrology (Amsterdam)*, *577*, 123,944–.

Behrangi, A., B. Khakbaz, J. A. Vrugt, Q. Duan, and S. Sorooshian (2008), Comment on "Dynamically dimensioned search algorithm for computationally efficient watershed model calibration" by Bryan A. Tolson and Christine A. Shoemaker, *Water Resources Research*, *44*(12).

Bentivoglio, R., E. Isufi, S. N. Jonkman, and R. Taormina (2022), Deep learning methods for flood mapping: a review of existing applications and future research directions, *Hydrology and Earth System Sciences*, *26*(16), 4345–4378, doi:10.5194/hess-26-4345-2022.

Bermúdez, M., L. Cea, and J. Puertas (2019), A rapid flood inundation model for hazard mapping based on least squares support vector machine regression, *Journal of Flood Risk Management*, *12*(S1), e12,522, doi:https://doi.org/10.1111/jfr3.12522.

Beven, K. (2005), On the concept of model structural error, *Water science and technology*, *52*(6), 167–175.

Bhupsingh, T., A. Graham, D. Peterson, S. Mennill, J. Brennan, and H. Collins (2022), Adapting to rising flood risk - an analysis of insurance solutions for canada.

Blum, C., and A. Roli (2003), Metaheuristics in combinatorial optimization: Overview and conceptual comparison, *ACM computing surveys*, *35*(3), 268–308.

Brocca, L., W. Zhao, and H. Lu (2023), High-resolution observations from space to address new applications in hydrology, *The Innovation*, *4*(3), 100,437, doi:https://doi.org/10.1016/j.xinn.2023.100437.

Brown, G., and J. R. Craig (2020), Structural calibration of an semi-distributed hydrological model of the liard river basin, *Canadian Water Resources Journal / Revue canadienne des ressources hydriques*, pp. 1–17, doi:10.1080/07011784.2020.1803143.

Brunner, G. (2022), Hec-ras, river analysis system hydraulic reference manual, *Tech. rep.*, US Army Corps of Engineers.

Bryant, S., H. McGrath, and M. Boudreault (2022), Gridded flood depth estimates from satellite-derived inundations, *Natural Hazards and Earth System Sciences*, *22*(4), 1437–1450, doi:10.5194/nhess-22-1437-2022.

Butts, M. B., J. T. Payne, M. Kristensen, and H. Madsen (2004), An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation, *Journal of Hydrology*, *298*(1-4), 242–266, doi:10.1016/j.jhydrol.2004.03.042.

Chadalawada, J., H. M. V. V. Herath, and V. Babovic (2020), Hydrologically informed machine learning for rainfall-runoff modeling: A genetic programming-based toolkit for automatic model induction, *Water Resources Research*, *56*(4).

Chang, L.-C., J.-Y. Liou, and F.-J. Chang (2022), Spatial-temporal flood inundation nowcasts by fusing machine learning methods and principal component analysis, *Journal of Hydrology*, *612*, 128,086, doi:https://doi.org/10.1016/j.jhydrol.2022.128086.

Chaudhuri, C., A. Gray, and C. Robertson (2021), Inundated-v1.0: A height above nearest drainage (hand)-based flood risk modeling system using a discrete global grid system, *Geoscientific Model Development*, *14*(6), 3295–3315.

Chicco, D. (2017), Ten quick tips for machine learning in computational biology, *BioData Mining*, *10*, doi:10.1186/s13040-017-0155-3.

Chlumsky, R., J. Mai, J. R. Craig, and B. A. Tolson (2021), Simultaneous calibration of hydrologic model structure and parameters using a blended model, *Water Resources Research*, *57*(5), e2020WR029,229, doi:https://doi.org/10.1029/2020WR029229.

Chlumsky, R., J. Mai, J. R. Craig, and B. A. Tolson (2024a), Advancement of a blended hydrologic model for robust model performance, *Journal of Hydrologic Engineering*, *29*(5), 04024,033, doi:10.1061/JHYEFF.HEENG-6246.

Chlumsky, R., J. R. Craig, and B. A. Tolson (2024b), A reach-integrated hydraulic modelling approach for large-scale and real-time inundation mapping, *Geoscientific Model Development Discussions*, *2024*, 1–28, doi:10.5194/gmd-2024-184.

Chow, V.-T. (1959), *Open-channel hydraulics*, McGraw-Hill civil engineering series, McGraw-Hill, New York.

Chu, H., W. Wu, Q. Wang, R. Nathan, and J. Wei (2020), An ann-based emulation modelling framework for flood inundation modelling: Application, challenges and future directions, *Environmental modelling & software : with environment data news*, *124*, 104,587–.

Cian, F., M. Marconcini, and P. Ceccato (2018), Normalized difference flood index for rapid flood mapping: Taking advantage of eo big data, *Remote Sensing of Environment*, *209*, 712–730, doi:https://doi.org/10.1016/j.rse.2018.03.006.

Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay (2008), Framework for understanding structural errors (fuse): A modular framework to diagnose differences between hydrological models, *Water Resources Research*, *44*(12), n/a–n/a, doi:10.1029/2007WR006735.

Clark, M. P., D. Kavetski, and F. Fenicia (2011), Pursuing the method of multiple working hypotheses for hydrological modeling, *Water Resources Research*, *47*(9), n/a–n/a, doi:10.1029/2010WR009827.

Clark, M. P., B. Nijssen, J. D. Lundquist, D. Kavetski, D. E. Rupp, R. A. Woods, J. E. Freer, E. D. Gutmann, A. W. Wood, L. D. Brekke, J. R. Arnold, D. J. Gochis, and R. M. Rasmussen (2015a), A unified approach for process-based hydrologic modeling: 1. Modeling concept, *Water Resources Research*, *51*(4), 2498–2514.

Clark, M. P., B. Nijssen, J. D. Lundquist, D. Kavetski, D. E. Rupp, R. A. Woods, J. E. Freer, E. D. Gutmann, A. W. Wood, D. J. Gochis, R. M. Rasmussen, D. G. Tarboton, V. Mahat, G. N. Flerchinger, and D. G. Marks (2015b), A unified approach for process-based hydrologic modeling: 2. Model implementation and case studies, *Water Resources Research*, *51*(4), 2515–2542, doi:https://doi.org/10.1002/2015WR017200.

Coron, L., V. Andréassian, C. Perrin, J. Lerat, J. Vaze, M. Bourqui, and F. Hendrickx (2012), Crash testing hydrological models in contrasted climate conditions: An experiment on 216 australian catchments, *Water Resources Research*, *48*(5), doi:10.1029/2011WR011721.

Craig, J. R. (2020), *Raven: User's and Developer's Manual v3.0*.

Craig, J. R., G. Brown, R. Chlumsky, R. W. Jenkinson, G. Jost, K. Lee, J. Mai, M. Serrer, N. Sgro, M. Shafii, A. P. Snowdon, and B. A. Tolson (2020), Flexible watershed simulation with the raven hydrological modelling framework, *Environmental Modelling and Software*, *129*, doi:10.1016/j.envsoft.2020.104728.

Cruise, J. F., M. M. M. M. Sherif, and V. P. V. P. Singh (2007), *Elementary hydraulics*, Thomson/Nelson, Toronto, Ont.

Cuntz, M., J. Mai, M. Zink, S. Thober, R. Kumar, D. Schäfer, M. Schrön, J. Craven, O. Rakovec, D. Spieler, V. Prykhodko, G. Dalmasso, J. Musuuza, B. Langenberg, s. Attinger, and L. Samaniego (2015), Computationally inexpensive identification of noninformative model parameters by sequential screening, *Water Resources Research*, *51*(8), 6417–6441.

Cuntz, M., J. Mai, L. Samaniego, M. Clark, V. Wulfmeyer, O. Branch, S. Attinger, and S. Thober (2016), The impact of standard and hard-coded parameters on the hydrologic fluxes in the noah-mp land surface model, *Journal of Geophysical Research: Atmospheres*, *121*(18), 10,676–10,700, doi:10.1002/2016JD025097.

Darbandsari, P., and P. Coulibaly (2019), Inter-Comparison of Different Bayesian Model Averaging Modifications in Streamflow Simulation, *Water*, *11*(8), 1707–27.

Diehl, R. M., J. D. Gourevitch, S. Drago, and B. C. Wemple (2021), Improving flood hazard datasets using a low-complexity, probabilistic floodplain mapping approach, *PLOS ONE*, *16*(3), 1–20, doi:10.1371/journal.pone.0248683.

Diks, C., and J. Vrugt (2010), Comparison of point forecast accuracy of model averaging methods in hydrologic applications, *Stochastic environmental research and risk assessment*, *24*(6), 809–820.

Duan, Q., S. Sorooshian, and V. Gupta (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water resources research*, *28*(4), 1015–1031.

Duan, Q., V. Gupta, and S. Sorooshian (1993), Shuffled complex evolution approach for effective and efficient global minimization, *Journal of optimization theory and applications*, *76*(3), 501–521.

Duan, Q., J. Schaake, V. Andréassian, S. Franks, G. Goteti, H. V. Gupta, Y. M. Gusev, F. Habets, A. Hall, L. Hay, T. Hogue, M. Huang, G. Leavesley, X. Liang, O. N. Nasonova, J. Noilhan, L. Oudin, S. Sorooshian, and T. Wagener (2006), Model parameter estimation experiment (mopex): An overview of science strategy and major results from the second and third workshops, *Journal of Hydrology*, *320*(1-2), 3–17, doi:10.1016/j.jhydrol.2005.07.031.

Duan, Q., N. K. Ajami, X. Gao, and S. Sorooshian (2007), Multi-model ensemble hydrologic prediction using Bayesian model averaging, *Advances in Water Resources*, *30*(5), 1371–1386.

Enriquez, A. R., T. Wahl, S. A. Talke, P. M. Orton, J. F. Booth, M. Agulles, and S. Santamaria-Aguilar (2023), Matflood: An efficient algorithm for mapping flood

extent and depth, *Environmental Modelling & Software*, *169*, 105,829, doi:https://doi.org/10.1016/j.envsoft.2023.105829.

Euser, T., H. C. Winsemius, M. Hrachowitz, F. Fenicia, S. Uhlenbrook, and H. H. G. Savenije (2013), A framework to assess the realism of model structures using hydrological signatures, *Hydrology and earth system sciences*, *17*(5), 1893–1912.

Evin, G., M. Thyer, D. Kavetski, D. Mcinerney, and G. Kuczera (2014), Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity, *Water Resources Research*, *50*(3), 2350–2375, doi:10.1002/2013WR014185.

Fenicia, F., H. H. G. Savenije, P. Matgen, and L. Pfister (2008), Understanding catchment behavior through stepwise model concept improvement, *Water Resources Research*, *44*(1), n/a–n/a, doi:10.1029/2006WR005563.

Fenicia, F., D. Kavetski, and H. H. G. Savenije (2011), Elements of a flexible approach for conceptual hydrological modeling: 1. motivation and theoretical development, *Water Resources Research*, *47*(11), n/a–n/a, doi:10.1029/2010WR010174.

Fenicia, F., D. Kavetski, H. H. G. Savenije, and L. Pfister (2016), From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions, *Water resources research*, *52*(2), 954–989.

Fohringer, J., D. Dransch, H. Kreibich, and K. Schröter (2015), Social media as an information source for rapid flood inundation mapping, *Natural Hazards and Earth System Sciences*, *15*(12), 2725–2738, doi:10.5194/nhess-15-2725-2015.

Follum, M. L. (2013), Autoroute rapid flood inundation model, *Tech. rep.*, Coastal and Hydraulics Laboratory (U.S.), Engineer Research and Development Center (U.S.).

Fraehr, N., Q. J. Wang, W. Wu, and R. Nathan (2023), Development of a fast and accurate hybrid model for floodplain inundation simulations, *Water Resources Research*, *59*(6), e2022WR033,836, doi:https://doi.org/10.1029/2022WR033836, e2022WR033836 2022WR033836.

Fraehr, N., Q. J. Wang, W. Wu, and R. Nathan (2024), Assessment of surrogate models for flood inundation: The physics-guided lsg model vs. state-of-the-art machine learning models, *Water Research*, *252*, 121,202, doi:https://doi.org/10.1016/j.watres.2024.121202.

Frame, J. M., F. Kratzert, D. Klotz, M. Gauch, G. Shalev, O. Gilon, L. M. Qualls, H. V. Gupta, and G. S. Nearing (2022), Deep learning rainfall–runoff predictions of extreme events, *Hydrology and Earth System Sciences*, *26*(13), 3377–3392, doi:10.5194/hess-26-3377-2022.

Frame, J. M., F. Kratzert, H. V. Gupta, P. Ullrich, and G. S. Nearing (2023), On strictly enforced mass conservation constraints for modelling the rainfall-runoff process, *Hydrological Processes*, *37*(3), e14,847, doi:https://doi.org/10.1002/hyp.14847.

Frame, J. M., T. Nair, V. Sunkara, P. Popien, S. Chakrabarti, T. Anderson, N. R. Leach, C. Doyle, M. Thomas, and B. Tellman (2024), Rapid inundation mapping using the us national water model, satellite observations, and a convolutional neural network, *Geophysical Research Letters*, *51*(17), e2024GL109,424, doi:https://doi.org/10.1029/2024GL109424.

Gavrilas, M. (2010), Heuristic and metaheuristic optimization techniques with application to power systems, in *Proceedings of the 12th WSEAS International Conference on Mathematical Methods and Computational Techniques in Electrical Engineering*, MMACTEE'10, p. 95–103, World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, USA.

Ghimire, E., S. Sharma, and N. Lamichhane (2022), Evaluation of one-dimensional and two-dimensional hec-ras models to predict flood travel time and inundation area for flood warning system, *ISH journal of hydraulic engineering*, *28*(1), 110–126.

Gilleland, E., and R. W. Katz (2016), extRemes 2.0: An extreme value analysis package in R, *Journal of Statistical Software*, *72*(8), 1–39, doi:10.18637/jss.v072.i08.

Glover, F. (1986), Future paths for integer programming and links to artificial intelligence, *Computers & Operations Research*, *13*(5), 533–549.

Golnaraghi, M., J. Thistlethwaite, D. Henstra, and C. Stewart (2020), Flood risk management in canada: Building flood resilience in a changing climate., *Tech. rep.*, The Geneva Association.

Government of Canada (2023), *Flood Hazard Identification and Mapping Program*, Government of Canada.

Guidolin, M., A. S. Chen, B. Ghimire, E. C. Keedwell, S. Djordjević, and D. A. Savić (2016), A weighted cellular automata 2d inundation model for rapid flood analysis, *Environmental Modelling & Software*, *84*, 378–394, doi:https://doi.org/10.1016/j.envsoft.2016.07.008.

Guillaume, J. H., J. D. Jakeman, S. Marsili-Libelli, M. Asher, P. Brunner, B. Croke, M. C. Hill, A. J. Jakeman, K. J. Keesman, S. Razavi, and J. D. Stigter (2019), Introductory overview of identifiability analysis: A guide to evaluating whether you have the right type of data for your modeling purpose, *Environmental Modelling and Software : with environment data news*, *119*, 418–432.

Gupta, H. V., T. Wagener, and Y. Liu (2008), Reconciling theory with observations: elements of a diagnostic approach to model evaluation, *Hydrological Processes*, *22*(18), 3802–3813, doi:10.1002/hyp.6989.

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

Gupta, H. V., M. P. Clark, J. A. Vrugt, G. Abramowitz, and M. Ye (2012), Towards a comprehensive assessment of model structural adequacy, *Water Resources Research*, *48*(8), n/a–n/a, doi:10.1029/2011WR011044.

Haddad, O. B., M. M. R. Tabari, E. Fallah-Mehdipour, and M. A. Mariño (2013), Groundwater model calibration by meta-heuristic algorithms, *Water resources management*, *27*(7), 2515–2529.

Hamidi, E., B. G. Peter, D. F. Muñoz, H. Moftakhari, and H. Moradkhani (2023), Fast flood extent monitoring with sar change detection using google earth engine, *IEEE Transactions on Geoscience and Remote Sensing*, *61*, 1–19, doi:10.1109/TGRS.2023.3240097.

Hashim, S., E. Suhana Mokhtar, A. Ikhwan Abdul Halim, W. Mohd Naim Wan Mohd, N. Aizam Adnan, and B. Pradhan (2021), Cross section intervals of flood intervals of flood inundation mapping at ungauged area, *IOP conference series. Earth and environmental science*, *620*(1), 12,003–.

Henstra, D., A. Minano, and J. Thistlethwaite (2019), Communicating disaster risk? an evaluation of the availability and quality of flood maps, *Natural Hazards and Earth System Sciences*, *19*(1), 313–323, doi:10.5194/nhess-19-313-2019.

Hoey, S. V., P. Seuntjens, J. van Der Kwast, and I. Nopens (2014), A qualitative model structure sensitivity analysis method to support model selection, *Journal of Hydrology*, *519*, 3426–3435, doi:10.1016/j.jhydrol.2014.09.052.

Hollander, M., and D. A. Wolfe (1973), *Nonparametric Statistical Methods*, 68-75 pp., John Wiley & Sons, New York.

Horton, P., B. Schaefli, and M. Kauzlaric (2021), Why do we have so many different hydrological models? a review based on the case of switzerland, *WIREs Water*, *n/a*(n/a), e1574, doi:10.1002/wat2.1574.

Hou, J., N. Zhou, G. Chen, M. Huang, and G. Bai (2021), Rapid forecasting of urban flood inundation using multiple machine learning models, *Natural hazards (Dordrecht)*, *108*(2), 2335–2356.

Huang, X., W. Liao, X. Lei, Y. Jia, Y. Wang, X. Wang, Y. Jiang, and H. Wang (2014), Parameter optimization of distributed hydrological model with a modified dynamically dimensioned search algorithm, *Environmental Modelling and Software : with environment data news*, *52*, 98–110.

Hulsing, H., W. Smith, and E. D. Cobb (1966), Velocity-head coefficients in open channels, *Tech. rep.*, United States Geoogical Survey.

Hydrologic Engineering Center (2023), *HEC-RAS User's Manual version 6.4.*

Jafarzadegan, K., H. Moradkhani, F. Pappenberger, H. Moftakhari, P. Bates, P. Abbaszadeh, R. Marsooli, C. Ferreira, H. L. Cloke, F. Ogden, and Q. Duan (2023), Recent advances and new frontiers in riverine and coastal flood modeling, *Reviews of Geophysics*, *61*(2), e2022RG000,788, doi:https://doi.org/10.1029/2022RG000788, e2022RG000788 2022RG000788.

Jamali, B., R. Löwe, P. M. Bach, C. Urich, K. Arnbjerg-Nielsen, and A. Deletic (2018), A rapid urban flood inundation and damage assessment model, *Journal of hydrology (Amsterdam)*, *564*, 1085–1098.

Jet Propulsion Laboratory (2024), Mapping earth's waterways from space: How swot transforms flood forecasting, *Tech. rep.*, SciTechDaily.

Kabir, S., S. Patidar, X. Xia, Q. Liang, J. Neal, and G. Pender (2020), A deep convolutional neural network model for rapid prediction of fluvial flood inundation, *Journal of Hydrology*, *590*, 125,481, doi:https://doi.org/10.1016/j.jhydrol.2020.125481.

Kavetski, D., and M. P. Clark (2010), Ancient numerical daemons of conceptual hydrological modeling: 2. impact of time stepping schemes on model analysis and prediction, *Water Resources Research*, *46*(10), n/a–n/a, doi:10.1029/2009WR008896.

Kavetski, D., and F. Fenicia (2011), Elements of a flexible approach for conceptual hydrological modeling: 2. application and experimental insights, *Water resources research*, *47*(11).

Kavetski, D., F. Fenicia, P. Reichert, and C. Albert (2018), Signature-domain calibration of hydrological models using approximate bayesian computation: Theory and comparison to existing applications, *Water resources research*, *54*(6), 4059–4083.

Klotz, D., F. Kratzert, M. Gauch, A. Keefe Sampson, J. Brandstetter, G. Klambauer, S. Hochreiter, and G. Nearing (2022), Uncertainty estimation with deep learning for rainfall–runoff modeling, *Hydrology and Earth System Sciences*, *26*(6), 1673–1693, doi:10.5194/hess-26-1673-2022.

Kneis, D. (2015), A lightweight framework for rapid development of object-based hydrological model engines, *Environmental Modelling and Software : with environment data news*, *68*, 110–121.

Knoben, W. J. M., J. E. Freer, K. J. A. Fowler, M. C. Peel, and R. A. Woods (2019), Modular assessment of rainfall–runoff models toolbox (marrmot) v1.2: an open-source, extendable framework providing implementations of 46 conceptual hydro-

logic models as continuous state-space formulations, *Geoscientific model development*, *12*(6), 2463–2480.

Knoben, W. J. M., J. E. Freer, M. C. Peel, K. J. A. Fowler, and R. A. Woods (2020), A brief analysis of conceptual model structure uncertainty using 36 models and 559 catchments, *Water Resources Research*, *56*(9), e2019WR025,975, doi:https://doi.org/10.1029/2019WR025975, e2019WR025975 10.1029/2019WR025975.

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

Kratzert, F., M. Herrnegger, D. Klotz, S. Hochreiter, and G. Klambauer (2019b), Neuralhydrology – interpreting lstms in hydrology, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, *11700*, 347–362.

Kreitmair, M. J. (2020), Derivation of the shallow water equations, in *The Effect of Uncertainty on Tidal Stream Energy Resource Estimates*, Springer Theses, pp. 17–25, Springer International Publishing AG, Switzerland.

Lane, R. A., G. Coxon, J. E. Freer, T. Wagener, P. J. Johnes, J. P. Bloomfield, S. Greene, C. J. A. Macleod, and S. M. Reaney (2019), Benchmarking the predictive capability of hydrological models for river flow and flood peak predictions across over 1000 catchments in great britain, *Hydrology and Earth System Sciences*, *23*(10), 4011–4032, doi:10.5194/hess-23-4011-2019.

Lane, S., K. Bradbrook, K. Richards, P. Biron, and A. Roy (1999), The application of computational fluid dynamics to natural river channels: three-dimensional versus two-dimensional approaches, *Geomorphology (Amsterdam)*, *29*(1), 1–20.

Leavesley, G. H., S. L. Markstrom, M. S. Brewer, and R. J. Viger (1996), The modular modeling system (mms) — the physical process modeling component of a database-centered decision support system for water and power management, *Water, Air, & Soil Pollution*, *90*(1), 303–311.

Leavesley, G. H., S. L. Markstrom, P. J. Restrepo, and R. J. Viger (2002), A modular approach to addressing model design, scale, and parameter estimation issues in distributed hydrological modelling, *Hydrological Processes*, *16*(2), 173–187.

Lees, T., S. Reece, F. Kratzert, D. Klotz, M. Gauch, J. De Bruijn, R. Kumar Sahu, P. Greve, L. Slater, and S. J. Dadson (2022), Hydrological concept formation inside long short-term memory (lstm) networks, *Hydrology and Earth System Sciences*, *26*(12), 3079–3101, doi:10.5194/hess-26-3079-2022.

Lhomme, J., P. Sayers, B. Gouldby, P. Samuels, M. Wills, and J. Mulet-Marti (2009a), Recent development and application of a rapid flood spreading method, in *Flood Risk Management: Research and Practice*, 1 ed., pp. 15–15, CRC Press, United Kingdom.

Lhomme, J., P. Sayers, B. Gouldby, P. Samuels, M. Wills, and J. Mulet-Marti (2009b), Recent development and application of a rapid flood spreading method, in *Flood Risk Management: Research and Practice*, 1 ed., pp. 15–15, CRC Press, United Kingdom.

Li, Z., C. Wang, C. T. Emrich, and D. Guo (2018), A novel approach to leveraging social media for rapid flood mapping: a case study of the 2015 south carolina floods, *Cartography and geographic information science*, *45*(2), 97–110.

Li, Z., F. Q. Duque, T. Grout, B. Bates, and I. Demir (2023), Comparative analysis of performance and mechanisms of flood inundation map generation using height above nearest drainage, *Environmental modelling & software : with environment data news*, *159*, 105,565–.

Lin, L., C. Tang, Q. Liang, Z. Wu, X. Wang, and S. Zhao (2023), Rapid urban flood risk mapping for data-scarce environments using social sensing and region-stable deep neural network, *Journal of Hydrology*, *617*, 128,758, doi:https://doi.org/10.1016/j.jhydrol.2022.128758.

Lindsay, J. B. (2016), The practice of dem stream burning revisited, *Earth Surface Processes and Landforms*, *41*(5), 658–668, doi:https://doi.org/10.1002/esp.3888.

Mai, J., J. R. Craig, and B. A. Tolson (2020), Simultaneously determining global sensitivities of model parameters and model structure, *Hydrology and Earth System Sciences*, *24*(12), 5835–5858, doi:10.5194/hess-24-5835-2020.

Mai, J., H. Shen, B. A. Tolson, É. Gaborit, R. Arsenault, J. R. Craig, V. Fortin, L. M. Fry, M. Gauch, D. Klotz, F. Kratzert, N. OBrien, D. G. Princz, S. R. Koya, T. Roy, F. Seglenieks, N. K. Shrestha, A. G. T. Temgoua, V. Vionnet, and J. W. Waddell (2022a), The Great Lakes Runoff Intercomparison Project Phase 4: the Great Lakes (GRIP-GL), *Hydrology and Earth System Sciences*, *26*, 3537–3572, doi:10.5194/hess-26-3537-2022.

Mai, J., J. R. Craig, B. A. Tolson, and R. Arsenault (2022b), The sensitivity of simulated streamflow to individual hydrologic processes across north america, *Nature communications*, *13*(1), 455–455, doi:10.1038/s41467-022-28010-7.

Mai, J., J. R. Craig, and B. A. Tolson (2022c), The pie sharing problem: Unbiased sampling of n+1 summative weights, *Environmental Modelling & Software*, *148*, 105,282, doi:10.1016/j.envsoft.2021.105282.

Maier, H., S. Razavi, Z. Kapelan, L. Matott, J. Kasprzyk, and B. Tolson (2019), Introductory overview: Optimization using evolutionary algorithms and other metaheuristics, *Environmental Modelling and Software : with environment data news*, *114*, 195–213.

Mangukiya, N. K., S. Kushwaha, and A. Sharma (2024), A novel multi-model ensemble framework for fluvial flood inundation mapping, *Environmental Modelling & Software*, *180*, 106,163, doi:https://doi.org/10.1016/j.envsoft.2024.106163.

Martel, J.-L., K. Demeester, F. Brissette, A. Poulin, and R. Arsenault (2017), HMETS—A Simple and Efficient Hydrology Model for Teaching Hydrological Modelling, Flow Forecasting and Climate Change Impacts, *International Journal of Engineering Education*, *33*(4), 1307–1316.

Matott, L. S. (2017), *OSTRICH – An Optimization Software Toolkit for Research Involving Computational Heuristics Documentation and User's Guide* , State University of New York at Buffalo Center for Computational Research, 17.12.19 ed.

McGrath, H., J.-F. Bourgon, J.-S. Proulx-Bourque, M. Nastev, and A. Abo El Ezz (2018), A comparison of simplified conceptual models for rapid web-based flood inundation mapping, *Natural hazards (Dordrecht)*, *93*(2), 905–920.

Md Ali, A., G. Di Baldassarre, and D. P. Solomatine (2015), Testing different cross-section spacing in 1d hydraulic modelling: a case study on johor river, malaysia, *Hydrological sciences journal*, *60*(2), 351–360.

Mizukami, N., M. P. Clark, A. J. Newman, A. W. Wood, E. D. Gutmann, B. Nijssen, O. Rakovec, and L. Samaniego (2017), Towards seamless large-domain parameter estimation for hydrologic models, *Water Resources Research*, *53*(9), 8020–8040.

Mizukami, N., O. Rakovec, A. J. Newman, M. P. Clark, A. W. Wood, H. V. Gupta, and R. Kumar (2019), On the choice of calibration metrics for "high-flow" estimation using hydrologic models, *Hydrology and Earth System Sciences*, *23*(6), 2601–2614, doi:10.5194/hess-23-2601-2019.

Moeini, A., B. Abbasi, and H. Mahlooji (2011), Conditional distribution inverse method in generating uniform random vectors over a simplex, *Communications in Statistics: Simulation and Computation*, *40*(5), 685–693, doi:10.1080/03610918.2010.551012.

Monaghan, J. (1994), Simulating free surface flows with sph, *Journal of computational physics*, *110*(2), 399–406.

Morvan, H., D. Knight, N. Wright, X. Tang, and A. Crossley (2008), The concept of roughness in fluvial hydraulics and its formulation in 1d, 2d and 3d numerical simulation models, *Journal of Hydraulic Research*, *46*, 191–208, doi:10.1080/00221686.2008.9521855.

Mugunthan, P., C. A. Shoemaker, and R. G. Regis (2005), Comparison of function approximation, heuristic, and derivative-based methods for automatic calibration of computationally expensive groundwater bioremediation models, *Water resources research*, *41*(11), W11,427–n/a.

Muhammad, A., T. Stadnyk, F. Unduche, and P. Coulibaly (2018), Multi-Model Approaches for Improving Seasonal Ensemble Streamflow Prediction Scheme with Various Statistical Post-Processing Techniques in the Canadian Prairie Region, *Water*, *10*(11), 1604–18.

Munasinghe, D., R. P. d. M. Frasson, C. H. David, M. Bonnema, G. Schumann, and G. R. Brakenridge (2023), A multi-sensor approach for increased measurements of floods and their societal impacts from space, *Communications Earth & Environment*, *4*, doi:10.1016/j.envsoft.2024.106163.

Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models part i — a discussion of principles, *Journal of Hydrology*, *10*(3), 282–290, doi:10.1016/0022-1694(70)90255-6.

Neal, J., G. Schumann, and P. Bates (2012), A subgrid channel model for simulating river hydraulics and floodplain inundation over large and data sparse areas, *Water Resources Research*, *48*(11), doi:https://doi.org/10.1029/2012WR012514.

Nearing, G. S., F. Kratzert, A. K. Sampson, C. S. Pelissier, D. Klotz, J. M. Frame, C. Prieto, and H. V. Gupta (2021), What role does hydrological science play in the age of machine learning?, *Water Resources Research*, *57*(3), e2020WR028,091, doi:https://doi.org/10.1029/2020WR028091, e2020WR028091 10.1029/2020WR028091.

Nelder, J. A., and R. Mead (1965), A simplex method for function minimization, *Computer journal*, *7*(4), 308–313.

Nevo, S., E. Morin, A. Gerzi Rosenthal, A. Metzger, C. Barshai, D. Weitzner, D. Voloshin, F. Kratzert, G. Elidan, G. Dror, G. Begelman, G. Nearing, G. Shalev, H. Noga, I. Shavitt, L. Yuklea, M. Royz, N. Giladi, N. Peled Levi, O. Reich, O. Gilon, R. Maor, S. Timnat, T. Shechter, V. Anisimov, Y. Gigi, Y. Levin, Z. Moshe, Z. Ben-Haim, A. Hassidim, and Y. Matias (2022), Flood forecasting with machine learning models in an operational framework, *Hydrology and Earth System Sciences*, *26*(15), 4013–4032, doi:10.5194/hess-26-4013-2022.

Nobre, A., L. Cuartas, M. Hodnett, C. Rennó, G. Rodrigues, A. Silveira, M. Waterloo, and S. Saleska (2011), Height above the nearest drainage – a hydrologically relevant new terrain model, *Journal of hydrology (Amsterdam)*, *404*(1), 13–29.

Néelz, S., and G. Pender (2010), Benchmarking of 2D Hydraulic Modelling Packages, in *delivering benefits through evidence*, Department for Environment Food and Rural Affairs.

Orellana, B., I. Pechlivanidis, N. Mcintyre, H. Wheater, and T. Wagener (2008), A toolbox for the identification of parsimonious semi-distributed rainfall-runoff models: Application to the upper lee catchment, in *International Congress on Environmental Modelling and Software*.

Perrin, C., C. Michel, and V. Andréassian (2003), Improvement of a parsimonious model for streamflow simulation, *Journal of Hydrology*, *279*, 275–289.

Pietroniro, A., D. L. Peters, D. Yang, J.-M. Fiset, R. Saint-Jean, V. Fortin, R. Leconte, J. Bergeron, G. L. Siles, M. Trudel, C. Garnaud, P. Matte, L. C. Smith, C. J. Gleason, and T. M. Pavelsky (2019), Canada's contributions to the swot mission – terrestrial hydrology(swot-c th), *Canadian Journal of Remote Sensing*, *45*(2), 116–138, doi: 10.1080/07038992.2019.1581056.

Pilz, T., T. Francke, G. Baroni, and A. Bronstert (2020), How to tailor my process-based hydrological model? dynamic identifiability analysis of flexible model structures, *Water resources research*.

Pomeroy, J. W., D. M. Gray, T. Brown, N. R. Hedstrom, W. L. Quinton, R. J. Granger, and S. K. Carey (2007), The cold regions hydrological model: a platform for basing process representation and model structure on physical evidence, *Hydrological processes*, *21*(19), 2650–2667.

Public Safety Canada (2024), Floods, accessed on August 7, 2024.

Qin, Y., D. Kavetski, and G. Kuczera (2018a), A robust gauss-newton algorithm for the optimization of hydrological models: From standard gauss-newton to robust gauss-newton, *Water Resources Research*, *54*(11), 9655–9683, doi:https://doi.org/10.1029/2017WR022488.

Qin, Y., D. Kavetski, and G. Kuczera (2018b), A robust gauss-newton algorithm for the optimization of hydrological models: Benchmarking against industry-standard algorithms, *Water resources research*, *54*(11), 9637–9654.

Rakovec, O., N. Mizukami, R. Kumar, A. J. Newman, S. Thober, A. W. Wood, M. P. Clark, and L. Samaniego (2019), Diagnostic Evaluation of Large-Domain Hydrologic Models Calibrated Across the Contiguous United States, *Journal of Geophysical Research: Atmospheres*, *124*(24), 13,991–14,007.

Raven Development Team (2024), *Raven: User's and Developer's Manual v3.8*.

Razavi, S., and H. V. Gupta (2015), What do we mean by sensitivity analysis? the need for comprehensive characterization of "global" sensitivity in earth and environmental systems models, *Water Resources Research*, *51*(5), 3070–3092, doi:https://doi.org/10.1002/2014WR016527.

Razavi, S., B. A. Tolson, and D. H. Burn (2012), Review of surrogate modeling in water resources, *Water Resources Research*, *48*(7), doi:https://doi.org/10.1029/2011WR011527.

Remmers, J. O., A. J. Teuling, and L. A. Melsen (2020), Can model structure families be inferred from model output?, *Environmental modelling and software : with environment data news*, pp. 104,817–.

Rennó, C. D., A. D. Nobre, L. A. Cuartas, J. V. Soares, M. G. Hodnett, J. Tomasella, and M. J. Waterloo (2008), Hand, a new terrain descriptor using srtm-dem: Mapping terra-firme rainforest environments in amazonia, *Remote Sensing of Environment*, *112*(9), 3469–3481, doi:https://doi.org/10.1016/j.rse.2008.03.018.

Roussel, J.-R., and D. Auty (2024), *Airborne LiDAR Data Manipulation and Visualization for Forestry Applications*, r package version 4.1.1.

Roussel, J.-R., D. Auty, N. C. Coops, P. Tompalski, T. R. Goodbody, A. S. Meador, J.-F. Bourdon, F. de Boissieu, and A. Achim (2020), lidr: An r package for analysis of airborne laser scanning (als) data, *Remote Sensing of Environment*, *251*, 112,061, doi:https://doi.org/10.1016/j.rse.2020.112061.

Sampson, C. C., A. M. Smith, P. D. Bates, J. C. Neal, L. Alfieri, and J. E. Freer (2015), A high-resolution global flood hazard model, *Water resources research*, *51*(9), 7358–7381.

Schaefli, B., and H. V. Gupta (2007), Do nash values have value?, *Hydrological processes*, *21*(15), 2075–2080.

Schumann, G. J.-P., K. M. Andreadis, and P. D. Bates (2014), Downscaling coarse grid hydrodynamic model simulations over large domains, *Journal of hydrology (Amsterdam)*, *508*, 289–298.

Scriven, B. W. G., H. McGrath, and E. Stefanakis (2021), Gis derived synthetic rating curves and hand model to support on-the-fly flood mapping, *Natural hazards (Dordrecht)*, *109*(2), 1629–1653.

Seiller, G., F. Anctil, and R. Roy (2017), Design and experimentation of an empirical multistructure framework for accurate, sharp and reliable hydrological ensembles, *Journal of hydrology (Amsterdam)*, *552*, 313–340.

Shaeri Karimi, S., N. Saintilan, L. Wen, and R. Valavi (2019), Application of machine learning to model wetland inundation patterns across a large semiarid floodplain, *Water Resources Research*, *55*(11), 8765–8778, doi:https://doi.org/10.1029/2019WR024884.

Shafii, M., and B. A. Tolson (2015a), Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives, *Water Resources Research*, *51*(5), 3796–3814, doi:10.1002/2014WR016520.

Shafii, M., and B. A. Tolson (2015b), Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives, *Water Resources Research*, *51*(5), 3796–3814, doi:10.1002/2014WR016520.

Shafii, M., N. Basu, J. R. Craig, S. L. Schiff, and P. V. Cappellen (2017), A diagnostic approach to constraining flow partitioning in hydrologic models using a multiobjective optimization framework, *Water Resources Research*, *53*(4), 3279–3301, doi:10.1002/2016WR019736.

Spieler, D., and N. Schütze (2024), Investigating the model hypothesis space: Benchmarking automatic model structure identification with a large model ensemble, *Water Resources Research*, *60*(7), e2023WR036,199, doi:https://doi.org/10.1029/2023WR036199, e2023WR036199 2023WR036199.

Spieler, D., J. Mai, J. R. Craig, B. A. Tolson, and N. Schütze (2020), Automatic model structure identification for conceptual hydrologic models, *Water resources research*, *56*(9).

Steinhausen, M., D. Paprotny, F. Dottori, N. Sairam, L. Mentaschi, L. Alfieri, S. Lüdtke, H. Kreibich, and K. Schröter (2022), Drivers of future fluvial flood risk change for residential buildings in europe, *Global Environmental Change*, *76*, 102,559, doi:10.1016/j.gloenvcha.2022.102559.

Stützle, T. (1998), Local search algorithms for combinatorial problems - analysis, improvements, and new applications, Ph.D. thesis, Technischen Universitat¨ Darmstadt.

Teng, J., J. Vaze, D. Dutta, and S. Marvanek (2015), Rapid inundation modelling in large floodplains using lidar dem, *Water resources management*, *29*(8), 2619–2636.

Teng, J., A. Jakeman, J. Vaze, B. Croke, D. Dutta, and S. Kim (2017), Flood inundation modelling: A review of methods, recent advances and uncertainty analysis, *Environmental modelling & software : with environment data news*, *90*, 201–216.

Teng, J., J. Vaze, S. Kim, D. Dutta, A. J. Jakeman, and B. F. W. Croke (2019), Enhancing the capability of a simple, computationally efficient, conceptual flood inundation model in hydrologically complex terrain, *Water resources management*, *33*(2), 831–845.

Thayer, J. B. (2017), Downstream regime relations for single-thread channels, *River Research and Applications*, *33*(1), 182–186, doi:https://doi.org/10.1002/rra.3053.

Thistlethwaite, J., D. Henstra, A. Minano, and S. Wilkinson (2017), Engaging canadians in flood risk management, *Tech. rep.*, Centre for International Governance Innovation.

Tolson, B. A., and C. A. Shoemaker (2007), Dynamically dimensioned search algorithm for computationally efficient watershed model calibration, *Water Resources Research*, *43*(1), doi:10.1029/2005WR004723.

Tolson, B. A., and C. A. Shoemaker (2008), Reply to comment on "Dynamically dimensioned search algorithm for computationally efficient watershed model calibration" by Ali Behrangi et al., *Water Resources Research*, *44*(12), n/a–n/a.

Tolson, B. A., M. A. Esfahani, A. C. Zecchin, and H. R. Maier (2008), A new algorithm for water distribution system optimization: Discrete dynamically dimensioned search, in *World Environmental and Water Resources Congress 2008*, pp. 1–10, doi: 10.1061/40976(316)494.

Tolson, B. A., M. Asadzadeh, H. R. Maier, and A. Zecchin (2009), Hybrid discrete dynamically dimensioned search (hd-dds) algorithm for water distribution system design optimization, *Water resources research*, *45*(12).

Tripathy, P., and T. Malladi (2022), Global flood mapper: a novel google earth engine application for rapid flood mapping using sentinel-1 sar, *Natural hazards (Dordrecht)*, *114*(2), 1341–1363.

van Meerveld, H. J. T., and J. J. McDonnell (2006), Threshold relations in subsurface stormflow: 2. the fill and spill hypothesis, *Water Resources Research*, *42*(2).

Velázquez, J. A., F. Anctil, and C. Perrin (2010), Performance and reliability of multimodel hydrological ensemble simulations based on seventeen lumped models and a thousand catchments, *Hydrology and earth system sciences*, *14*(11), 2303–2317.

Wagener, T., and H. V. Gupta (2005), Model identification for hydrological forecasting under uncertainty, *Stochastic Environmental Research and Risk Assessment*, *19*(6), 378–387.

Wagener, T., M. Lees, H. Wheater, V. Singh, and D. Frevert (2002), A toolkit for the development and application of parsimonious hydrological models., in *Mathematical models of small watershed hydrology - Volume 2*, Water Resources Publications LLC, USA.

Wagener, T., N. McIntyre, M. J. Lees, H. S. Wheater, and H. V. Gupta (2003), Towards reduced uncertainty in conceptual rainfall-runoff modelling: dynamic identifiability analysis, *Hydrological Processes*, *17*(2), 455–476.

Weiler, M., and K. Beven (2015), Do we need a community hydrological model?, *Water Resources Research*, *51*(9), 7777–7784, doi:https://doi.org/10.1002/2014WR016731.

Wijaya, O. T., T.-H. Yang, H.-M. Hsu, and P. Gourbesville (2023), A rapid flood inundation model for urban flood analyses, *MethodsX*, *10*, 102,202, doi:https://doi.org/10.1016/j.mex.2023.102202.

Wilkerson, G. V., and G. Parker (2011), Physical basis for quasi-universal relationships describing bankfull hydraulic geometry of sand-bed rivers, *Journal of Hydraulic Engineering*, *137*(7), 739–753, doi:10.1061/(ASCE)HY.1943-7900.0000352.

Wing, O. E. J., P. D. Bates, C. C. Sampson, A. M. Smith, K. A. Johnson, and T. A. Erickson (2017), Validation of a 30 m resolution flood hazard model of the conterminous united states, *Water Resources Research*, *53*(9), 7968–7986, doi:https://doi.org/10.1002/2017WR020917.

Wu, Y., J. Liu, and D. Li (2020), Chaotic dynamically dimensioned search algorithm, *IEEE access*, *8*, 152,474–152,499.

Yang, J., R. D. Townsend, and B. Daneshfar (2006), Applying the hec-ras model and gis techniques in river network floodplain delineation, *Canadian journal of civil engineering*, *33*(1), 19–28.

Ye, A., Q. Duan, X. Yuan, E. F. Wood, and J. Schaake (2014), Hydrologic post-processing of mopex streamflow simulations, *Journal of Hydrology*, *508*, 147.

Yen, H., J. Jeong, W.-H. Tseng, M.-K. Kim, R. M. Records, and M. Arabi (2015), Computational procedure for evaluating sampling techniques on watershed model calibration, *Journal of Hydrologic Engineering*, *20*(7).

Zheng, X., D. G. Tarboton, D. R. Maidment, Y. Y. Liu, and P. Passalacqua (2018a), River channel geometry and rating curve estimation using height above the nearest drainage, *Journal of the American Water Resources Association*, *54*(4), 785–806.

Zheng, X., D. R. Maidment, D. G. Tarboton, Y. Y. Liu, and P. Passalacqua (2018b), Geoflood: Large-scale flood inundation mapping based on high-resolution terrain analysis, *Water Resources Research*, *54*(12), 10,013–10,033, doi:https://doi.org/10.1029/2018WR023457.

Zhou, Y., W. Wu, R. Nathan, and Q. J. Wang (2022), Deep learning-based rapid flood inundation modeling for flat floodplains with complex flow paths, *Water Resources Research*, *58*(12), e2022WR033,214, doi:https://doi.org/10.1029/2022WR033214, e2022WR033214 2022WR033214.

Zink, M., J. Mai, M. Cuntz, and L. Samaniego (2018), Conditioning a hydrologic model using patterns of remotely sensed land surface temperature, *Water Resources Research*, *54*(4), 2976–2998.

Appendix

Table 7.1 Processes and process options used for the Raven setup. In total $3 \times 3 \times 2 \times 2 \times 3 \times 1 \times 1 \times 1 \times 1 \times 1 \times 1 \times 1 = 108$ models are possible. The first option of each process M_1, N_1, O_1, P_1, Q_1, R_1, S_1, T_1, U_1, V_1, W_1, and X_1 resemble the HMETS model. All other combinations are artificial models. All process options, however, are used in different hydrologic models. The model parameters active in each option are listed as well. The ranges and a description of the parameters can be found in Table 7.2. This table is adapted from that of *Mai et al.* (2020).

Process	**Process option**		**Parameters active**
Processes with multiple options:			
Infiltration	M_1	INF_HMETS	$\{x_1, x_{29}\}$
”	M_2	INF_VIC_ARNO	$\{x_2, x_{29}\}$
”	M_3	INF_HBV	$\{x_3, x_{29}\}$
Quickflow	N_1	BASE_LINEAR_ANALYTIC	$\{x_4, x_{29}\}$
”	N_2	BASE_VIC	$\{x_5, x_6, x_{29}\}$
”	N_3	BASE_TOPMODEL	$\{x_5, x_6, x_7, x_{29}\}$
Soil evaporation	O_1	SOILEVAP_ALL	$\{x_8, x_{29}\}$
”	O_2	SOILEVAP_TOPMODEL	$\{x_8, x_9, x_{10}, x_{29}\}$
Baseflow	P_1	BASE_LINEAR_ANALYTIC	$\{x_{11}\}$
”	P_2	BASE_POWER_LAW	$\{x_{11}, x_{12}\}$
Snow balance	Q_1	SNOBAL_HMETS	$\{x_{13}, \ldots, x_{18}\}$
”	Q_2	SNOBAL_SIMPLE_MELT	–
”	Q_3	SNOBAL_HBV	$\{x_{18}, x_{19}\}$
Processes with single option:			
Convolution (surface runoff)	R_1	CONVOL_GAMMA	$\{x_{20}, x_{21}\}$
Convolution (delayed runoff)	S_1	CONVOL_GAMMA_2	$\{x_{22}, x_{23}\}$
Potential melt	T_1	POTMELT_HMETS	$\{x_{24}, x_{25}, x_{26}, x_{27}\}$
Percolation	U_1	PERC_LINEAR	$\{x_{28}, x_{29}, x_{30}\}$
Rain-snow partitioning	V_1	RAINSNOW_HBV	$\{x_{31}, x_{32}\}$
Precipitation correction	W_1	RAINSNOW_CORRECTION	$\{x_{33}, x_{34}\}$
Processes with single option but no tunable parameter combined to process X_1:			
Extraterr. Shortwave Gener.	X_1	SW_RAD_NONE	–
Potential evapotranspiration	X_1	PET_DATA	–
In-catchment routing	X_1	ROUTE_DUMP	–
In-channel routing	X_1	ROUTE_NONE	–

Table 7.2 The model parameters x_i used for the Raven setup. The parameter ranges used in calibration are provided. The process option shows where the corresponding parameter is active. The Raven table and parameter name can be used to locate the parameter in the Raven setup files. The TOPSOIL is the upper soil layer while PHREATIC is the lower soil layer. The three Raven parameters FIELD_CAPACITY TOPSOIL, SNOW_SWI_MAX, and MAX_MELT_FACTOR are derived using a sampled parameter (x_{10}, x_{14}, and x_{25}) and SAT_WILT TOPSOIL, SNOW_SWI_MIN, and MIN_MELT_FACTOR, respectively, to make sure that one parameter is always larger than the other. The baseflow coefficients BASEFLOW_COEFF TOPSOIL and PHREATIC are derived from parameters x_4 and x_{11} to allow for a logarithmic sampling. The weight-generating parameters are parameters x_{35} to x_{42} with a range of $[0 \dots 1]$, and are omitted from the table. This table is adapted from (*Mai et al.*, 2020).

Param.	**Range**	**Unit**	**Proc. Opt.**	**Raven table**	**Parameter name**
Infiltration:					
x_1	$[0.0, 1.0]$	-	M_1	LandUseParameterList	HMETS_RUNOFF_COEFF
x_2	$[0.1, 3.0]$	-	M_2	SoilParameterList	B_EXP TOPSOIL
x_3	$[0.5, 3.0]$	-	M_3	SoilParameterList	HBV_BETA TOPSOIL
Quickflow:					
x_4	$[-5.0, -1.0]$	1/d	N_1	SoilParameterList	BASEFLOW_COEFF TOPSOIL = 10.0^{x_4}
x_5	$[0.0, 100.0]$	mm/d	N_2, N_3	SoilParameterList	MAX_BASEFLOW_RATE TOPSOIL
x_6	$[0.5, 2.0]$	-	N_2, N_3	SoilParameterList	BASEFLOW_N TOPSOIL
x_7	$[5.0, 10.0]$	m	N_3	TerrainClasses	TOPMODEL_LAMBDA
Evaporation:					
x_8	$[0.0, 3.0]$	-	O_1, O_2	SoilParameterList	PET_CORRECTION TOPSOIL
x_9	$[0.0, 0.05]$	frac	O_2	SoilParameterList	SAT_WILT TOPSOIL
x_{10}	$[0.0, 0.45]$	frac	O_2	SoilParameterList	FIELD_CAPACITY TOPSOIL = SAT_WILT TOPSOIL + x_{10}
Baseflow:					
x_{11}	$[-5.0, -2.0]$	1/d	P_1, P_2	SoilParameterList	BASEFLOW_COEFF PHREATIC = $10.0^{x_{11}}$
x_{12}	$[0.5, 2.0]$	-	P_2	SoilParameterList	BASEFLOW_N PHREATIC
Snow balance:					
x_{13}	$[0.0, 0.1]$	frac	Q_1	GlobalParameter	SNOW_SWI_MIN
x_{14}	$[0.01, 0.3]$	frac	Q_1	GlobalParameter	SNOW_SWI_MAX = SNOW_SWI_MIN + x_{14}
x_{15}	$[0.005, 0.1]$	1/mm	Q_1	GlobalParameter	SWI_REDUCT_COEFF
x_{16}	$[-5.0, 2.0]$	°C	Q_1	LandUseParameterList	DD_REFREEZE_TEMP
x_{17}	$[0.0, 1.0]$	-	Q_1	LandUseParameterList	REFREEZE_EXP
x_{18}	$[0.0, 5.0]$	mm/d/°C	Q_1, Q_3	LandUseParameterList	REFREEZE_FACTOR
x_{19}	$[0.0, 0.4]$	frac	Q_3	GlobalParameter	SNOW_SWI
Convolution (surface runoff):					
x_{20}	$[0.3, 20.0]$	-	R_1	LandUseParameterList	GAMMA_SHAPE
x_{21}	$[0.01, 5.0]$	-	R_1	LandUseParameterList	GAMMA_SCALE
Convolution (delayed runoff):					
x_{22}	$[0.5, 13.0]$	-	S_1	LandUseParameterList	GAMMA_SHAPE2
x_{23}	$[0.15, 1.5]$	-	S_1	LandUseParameterList	GAMMA_SCALE2
Potential melt:					
x_{24}	$[1.5, 3.0]$	mm/d/°C	T_1	LandUseParameterList	MIN_MELT_FACTOR
x_{25}	$[0.0, 5.0]$	mm/d/°C	T_1	LandUseParameterList	MAX_MELT_FACTOR =

Continued on next page

Table 7.2 – *Continued from previous page*

Param.	**Range**	**Unit**	**Proc. Opt.**	**Raven table**	**Parameter name**
					MIN_MELT_FACTOR + x_{25}
x_{26}	$[-1.0, 1.0]$	°C	T_1	LandUseParameterList	DD_MELT_TEMP
x_{27}	$[0.01, 0.2]$	1/mm	T_1	LandUseParameterList	DD_AGGRADATION
Percolation:					
x_{28}	$[0.00001, 0.02]$	1/d	U_1	SoilParameterList	PERC_COEFF TOPSOIL
Rain-snow partitioning:					
x_{31}	$[-3.0, 3.0]$	°C	V_1	GlobalParameter	RAINSNOW_TEMP
x_{32}	$[0.5, 4.0]$	°C	V_1	GlobalParameter	RAINSNOW_DELTA
Precipitation correction:					
x_{33}	$[0.8, 1.2]$	-	W_1	Gauge	RAINCORRECTION
x_{34}	$[0.8, 1.2]$	-	W_1	Gauge	SNOWCORRECTION
Soil model:					
x_{29}	$[0.0, 0.5]$	m	$M_{1,2,3}$,$N_{1,2,3}$ $O_{1,2,3,4}$, U_1	SoilProfiles	thickness TOPSOIL
x_{30}	$[0.0, 2.0]$	m	U_1	SoilProfiles	thickness PHREATIC

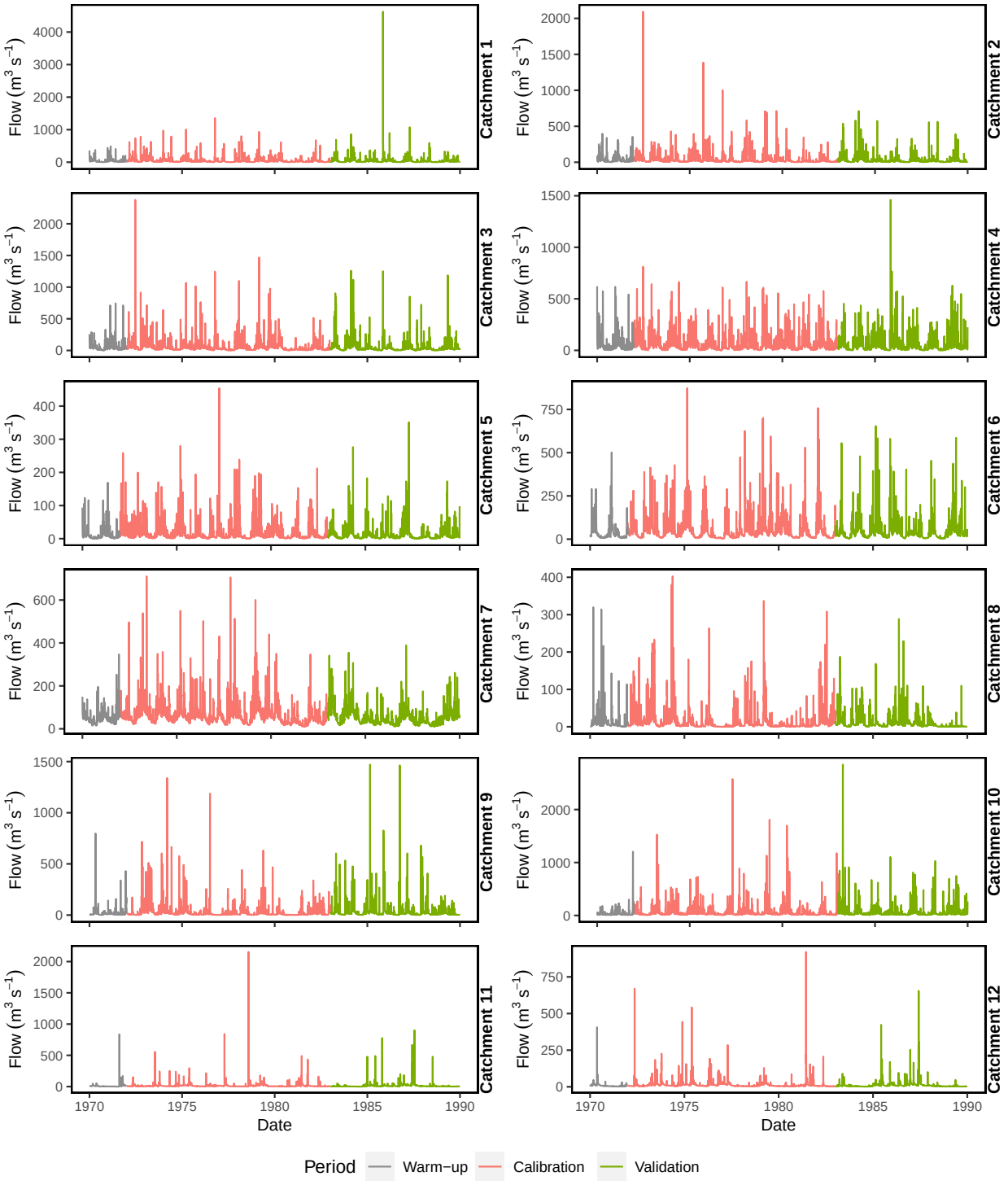


Figure 7.1 Observed hydrographs in each catchment by period, including the two year warm up period from January 1970 to December 1971 (grey lines), the twelve year calibration period from January 1972 to December 1983 (red lines), and the six year validation period from January 1984 to December 1989 (green lines). The hydrograph is provided for all twelve MOPEX catchments.

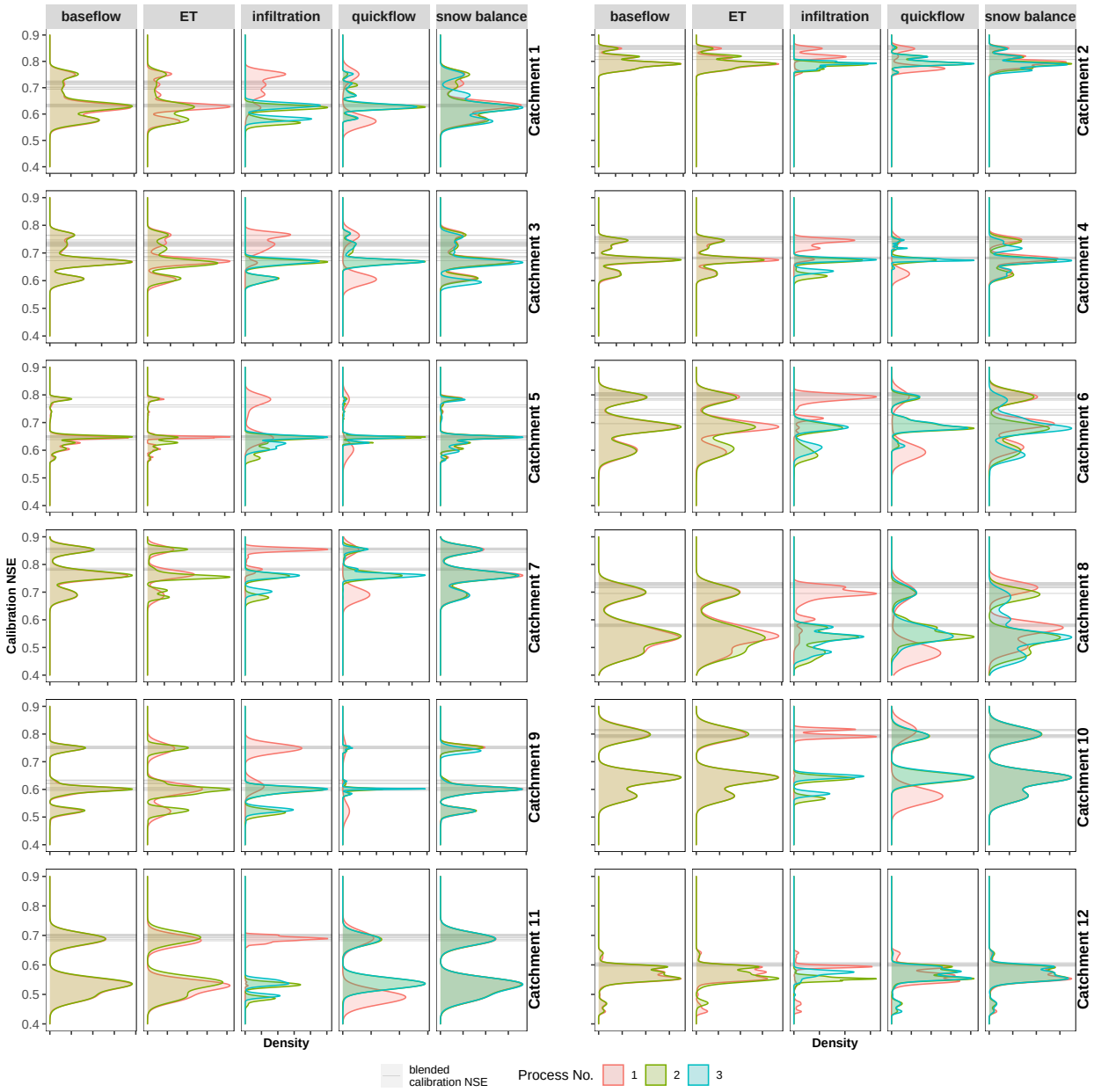


Figure 7.2 Model performance in calibration NSE score on the y-axis and density on the x-axis (1-109), plotted for catchments 1, 8, and 11, and faceted horizontally by the five groups of process algorithms in this study. Density axis ranges vary by plot but labels have been removed for simplicity. In each density curve the colour corresponds to a choice of process algorithm within the process group as 1, 2, or 3. Model structures 1 through 108 are fixed model structures. The horizontal grey lines indicate the NSE performance of the blended model trials. The averaged models are omitted from this analysis. Results are shown for a calibration budget of $10k$ model runs, and all twelve MOPEX catchments.

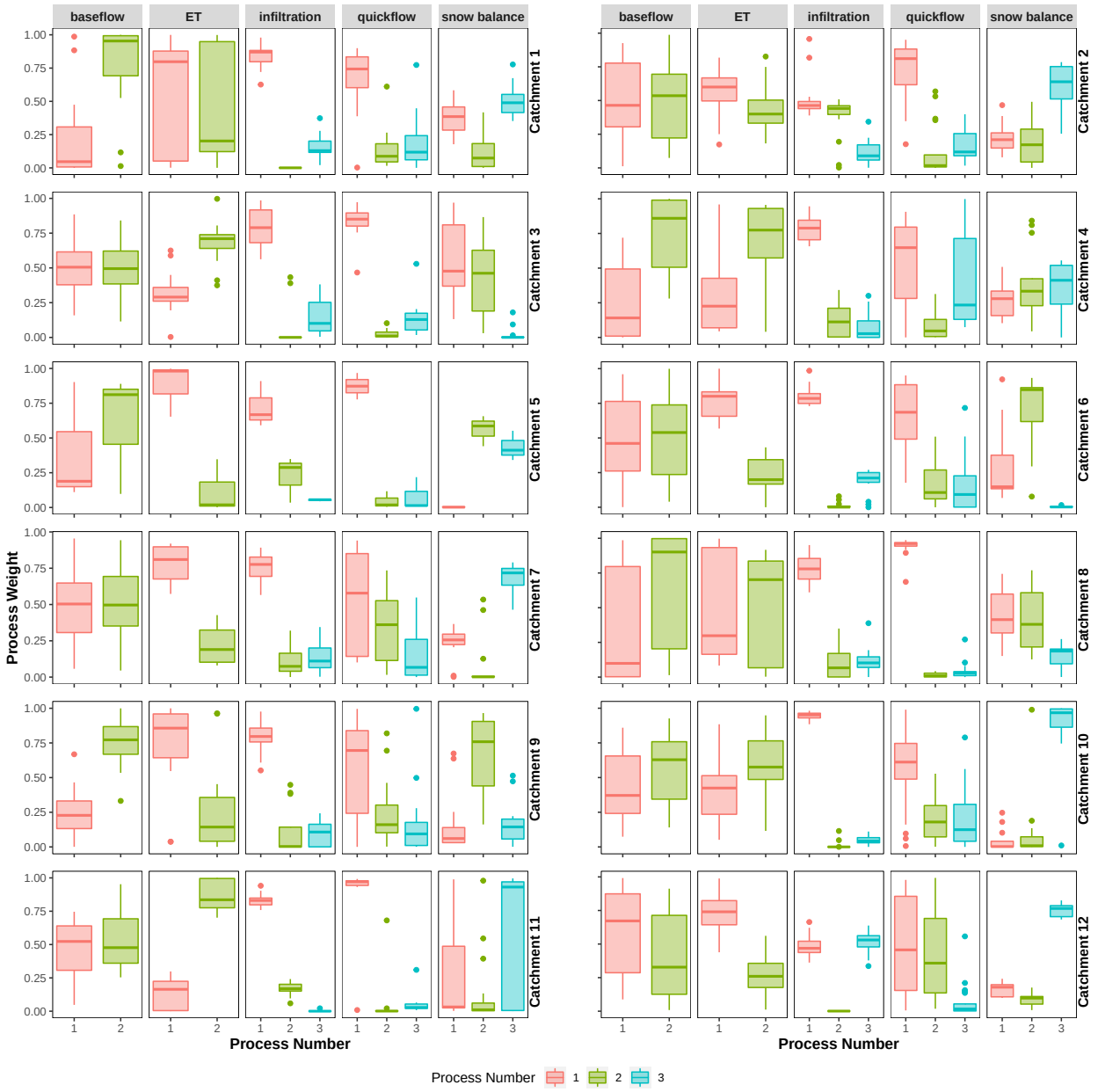


Figure 7.3 Blended model process weights from 0 to 1 on the y-axis, and numbered options for each blended hydrologic process on the x-axis, faceted by the hydrologic process group. Each box and whisker plot displays the spread of the given process weight within the process group, where horizontal lines represent the 25^{th}, 50^{th}, and 75^{th} percentiles, and outliers are shown as points. For each catchment, the data blended model result is filtered to include only information from independent trials where the calibration NSE performance was within 0.05 inclusive of the top performing independent trial. Results are shown for a calibration budget of $10k$ model runs, and all twelve MOPEX catchments.

Table 7.3 Summary information for the MOPEX catchments as well as the 24 independent catchments selected from HYSETS for validating the blended model version 2. PET and AET are model derived estimates as this data was not available for all stations, precipitation and streamflow data are derived from historical data. All hydrological metrics are determined for the 1970-1989 period in which the models are run in this study.

Catchment Index	Basin ID	State	Area (km^2)	Elevation (m)	Annual Avg. Precip (mm)	Annual Avg. PET (mm)	Annual Avg. Flow (mm)	Runoff Ratio
1	01608500	WV	3782	661	935	773	408	0.44
2	01643000	MD	2115	195	1059	749	458	0.43
3	01668000	VA	4129	207	1078	782	396	0.37
4	03054500	WV	2366	719	1318	754	774	0.59
5	03179000	WV	1023	793	984	557	441	0.45
6	03364000	IN	4419	269	1032	496	372	0.36
7	03451500	NC	2446	830	1510	771	764	0.51
8	05455500	IA	1486	253	867	764	256	0.29
9	07186000	MO	3013	332	1079	695	315	0.29
10	07378500	LA	3314	82	1575	1294	617	0.39
11	08167500	TX	3404	544	772	482	127	0.16
12	08172000	TX	2169	299	844	527	178	0.21
13	01017000	ME	4282	275	986	586	582	0.59
14	01122500	CT	1046	195	1216	595	685	0.56
15	01318500	NY	4308	529	1103	546	659	0.60
16	02472000	MS	1924	125	1438	1219	568	0.39
17	03434500	TN	1768	237	1357	615	567	0.42
18	10150500	UT	1688	2170	532	329	145	0.27
19	11189500	CA	1372	2358	377	506	97	0.26
20	11272500	CA	3296	1383	891	891	187	0.21
21	11276500	CA	1183	2713	1136	1014	299	0.26
22	13073000	ID	1491	1844	521	430	142	0.27
23	13139500	ID	1657	2348	707	412	272	0.39
24	13200000	ID	1028	1552	819	827	270	0.33
25	02056000	VA	1318	536	1043	579	373	0.36
26	02225500	GA	2874	86	1164	1088	316	0.27
27	02322500	FL	2633	39	1330	252	530	0.40
28	03249500	KY	2141	322	1195	1004	457	0.38
29	04208000	OH	1830	333	995	629	492	0.49
30	05356000	WI	2045	448	838	649	320	0.38
31	05594800	IL	1201	159	1000	818	260	0.26
32	06890900	KS	2954	339	917	914	215	0.23
33	13235000	ID	1155	2084	956	381	689	0.72
34	07069500	AR	3055	230	1166	1013	421	0.36
35	07147070	KS	1103	429	847	566	137	0.16
36	13292000	OR	1610	1562	874	348	298	0.34

Table 7.4 Process options included within each blended model configuration. Model configurations 25-28 & 31-32 were used for testing weighting schemes and are omitted. It is noted that this table displays blended options only; other structural changes are described in Section 2.2 by model configuration, and subsequent model configurations may not be the same as or subsets of earlier configurations depending on other changes that may have been implemented beyond what is shown in this table.

	AET					**BASE-FLOW**				**INFIL-TRATION**					**PET**												**POT-MELT**				**QUICK-FLOW**				**SNOW-BALANCE**				
Model Configuration	SOILEVAP_ALL	SOILEVAP_LINEAR	SOILEVAP_ROOT	SOILEVAP_SEQUEN	SOILEVAP_TOPMODEL	BASE_LINEAR_ANALYTIC	BASE_POWER_LAW	BASE_THRESH_POWER	BASE_VIC	INF_GA_SIMPLE	INF_GR4J	INF_HBV	INF_HMETS	INF_PARTITION	INF_VIC_ARNO	PET_DATA	PET_GRANGERGRAY	PET_HAMON	PET_HARGREAVES_1985	PET_LINACRE	PET_OUDIN	PET_PENMAN_MONTEITH	PET_PENMAN_SIMPLE33	PET_PENMAN_SIMPLE39	PET_PRIESTLEY_TAYLOR	POTMELT_CRHM_EBSM	POTMELT_EB	POTMELT_HMETS	POTMELT_RESTRICTED	POTMELT_USACE	BASE_LINEAR_ANALYTIC	BASE_POWER_LAW	BASE_THRESH_POWER	BASE_TOPMODEL	BASE_VIC	SNOBAL_CRHM_EBSM	SNOBAL_HBV	SNOBAL_HMETS	SNOBAL_SIMPLE_MELT	Number of Model Parameters
1. Original model version																																						
0	X				X	X	X					X	X		X						X						X				X			X	X	X			43
2. Original model version but using data sourced from MOPEX/HYSETS rather than PET_OUDIN																																						
1	X				X	X	X					X	X		X	X											X				X			X	X	X			43
3. Initial update based on expert consideration of process options																																						
2	X				X		X	X		X		X			X	X											X				X		X		X	X			45
4. Addition of depression storage & seepage, followed by adjustments to other process options																																						
3	X				X		X	X		X		X			X	X											X				X		X		X	X			48
4	X			X	X		X	X		X		X			X	X											X				X		X		X	X			49
5	X		X		X		X	X		X		X			X	X											X				X		X		X	X			51
5. Introduction of blended forcings, followed by iteration of process options																																						
6	X		X		X		X	X		X		X			X					X		X	X				X				X		X		X	X			54
7	X		X		X		X	X		X		X			X					X		X	X				X				X		X		X	X			57
8	X		X		X		X	X		X		X		X					X		X		X			X		X			X		X		X	X			55
9	X		X		X		X	X		X		X		X			X		X		X					X		X			X		X		X	X			58
10	X		X		X		X	X		X		X	X				X		X		X					X		X			X		X		X	X			59
11	X		X		X		X	X		X		X						X		X			X			X			X		X		X		X	X			56
12	X		X		X		X	X		X		X					X							X	X			X			X		X		X	X			59
13	X		X		X		X	X		X		X					X			X			X				X	X			X		X		X	X			58
14	X		X		X		X	X		X		X					X		X		X					X		X			X		X		X	X			58
15	X		X		X		X	X		X		X					X			X		X					X		X			X		X		X	X			55
16	X		X		X		X	X		X		X					X				X		X				X		X			X		X		X	X			55
17	X			X	X		X	X		X		X					X				X			X			X		X			X		X		X	X			52
18	X			X	X		X	X		X		X					X				X			X				X		X			X		X		X	X			44
19	X		X		X		X	X		X		X					X				X			X					X	X	X		X		X		X	X			50
20	X		X		X		X	X		X		X					X				X			X					X	X	X		X		X		X	X			50
21	X		X		X		X	X		X		X					X				X			X			X			X	X		X		X		X	X			50
22	X		X		X		X	X		X		X					X				X			X					X	X	X		X		X		X	X			50
23	X			X	X		X	X		X		X					X				X			X			X		X			X		X		X	X			47
6. Conglomerate model configuration																																						
24	X		X	X	X	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	79
7. Experiments in reducing complexity of the blended configuration																																						
29	X			X	X		X	X				X					X	X				X					X				X		X				X				41
30	X			X	X		X	X				X					X	X				X					X				X		X				X				40
33	X			X	X			X				X					X	X				X					X						X				X				37
8. Addition of non-blended processes, and additional complexity reduction experiments																																						
34	X			X	X			X				X					X	X				X						X					X				X				45
35	X			X	X		X	X				X					X	X				X						X					X				X				51
36	X			X	X		X	X				X					X	X				X						X					X				X				48
37	X			X	X			X				X					X	X				X						X					X				X				44

Table 7.5 Processes and process options used for the blended model version 2 setup (model configuration 36) in Raven. The model parameters active in each option are listed as well. The ranges and a description of the parameters can be found in Table 7.6. Note that weight-generating parameters used to determine process weights within each process group are not listed in this table.

Process	**Process option**		**Parameters active**
Processes with multiple (blended) options:			
Quickflow	H_1	BASE_POWER_LAW	$\{x_4, x_6, x_{29}\}$
”	H_2	BASE_THRESH_POWER	$\{x_5, x_6, x_{38}, x_{29}\}$
Soil evaporation (AET)	I_1	SOILEVAP_ALL	$\{x_8, x_{29}\}$
”	I_2	SOILEVAP_ROOT	$\{x_8, x_9, x_{10}, x_{29}\}$
”	I_2	SOILEVAP_SEQUEN	$\{x_8, x_9, x_{10}, x_{29}\}$
Baseflow	J_1	BASE_POWER_LAW	$\{x_{11}, x_{12}, x_{30}\}$
”	J_2	BASE_THRESH_POWER	$\{x_{44}, x_{12}, x_{39}, x_{30}\}$
PET	K_1	PET_GRANGERGRAY	–
”	K_2	PET_HAMON	–
”	K_3	PET_PENMAN_MONTEITH	$\{x_{55}\}$
POTMELT	L_1	POTMELT_HMETS	$\{x_{24}, x_{25}, x_{26}, x_{27}\}$
”	L_2	POTMELT_RESTRICTED	–
Processes with single option:			
Infiltration	M_1	INF_HMETS	$\{x_1, x_{29}\}$
Snow balance	N_1	SNOBAL_HBV	$\{x_{18}, x_{19}\}$
Canopy interception	O_1	PRECIP_ICEPT_USER	$\{x_{49}, x_{52}, x_{55}, x_{56}\}$
Canopy drip	P_1	CANDRIP_RUTTER	$\{x_{52}, x_{55}, x_{56}\}$
Canopy evaporation	Q_1	CANEVP_MAXIMUM	$\{x_8, x_{48}, x_{49}, x_{52}, x_{55}\}$
Abstraction	R_1	ABST_PERCENTAGE	$\{x_{41}\}$
Depression seepage	S_1	SEEP_LINEAR	$\{x_{40}\}$
Groundwater upwelling	T_1	CRISE_HBV	$\{x_{53}, x_{54}\}$
Percolation	U_1	PERC_LINEAR	$\{x_{28}, x_{29}, x_{35}, x_{30}\}$
Convolution (surface runoff)	V_1	CONVOL_GAMMA	$\{x_{20}, x_{21}\}$
Convolution (delayed runoff)	W_1	CONVOL_GAMMA_2	$\{x_{22}, x_{23}\}$
Rain-snow partitioning	X_1	RAINSNOW_HBV	$\{x_{31}, x_{32}\}$
Precipitation correction	Y_1	RAINSNOW_CORRECTION	$\{x_{33}, x_{34}\}$
Processes with single option but no tunable parameter combined to process:			
Extraterr. Shortwave Gener.	Z_1	SW_RAD_DEFAULT	–
In-catchment routing	Z_2	ROUTE_DUMP	–
In-channel routing	Z_3	ROUTE_NONE	–

Table 7.6 The model parameters x_i used for the blended model version 2 setup (model configuration 36) in Raven. The parameter numbering is discontinuous as not all parameters across model configurations are used in model configuration 36. The parameter ranges used in calibration are provided. The process option shows where the corresponding parameter is active. The Raven table and parameter name can be used to locate the parameter in the Raven setup files. The TOPSOIL is the upper soil layer while PHREATIC is the lower soil layer. The three Raven parameters FIELD_CAPACITY TOPSOIL, SNOW_SWI_MAX, and MAX_MELT_FACTOR are derived using a sampled parameter (x_{10}, x_{14}, and x_{25}) and SAT_WILT TOPSOIL, SNOW_SWI_MIN, and MIN_MELT_FACTOR, respectively, to make sure that one parameter is always larger than the other. The baseflow coefficients BASEFLOW_COEFF TOPSOIL and PHREATIC are derived from parameters x_4 and x_{11} to allow for a logarithmic sampling. The weight-generating parameters, with a range of $[0 \dots 1]$, are omitted from the table.

Param.	Range	Unit	Proc. Opt.	Raven table	Parameter name
Quickflow:					
x_4	$[-5.0, -1.0]$	1/d	H_1	SoilParameterList	BASEFLOW_COEFF TOPSOIL = 10.0^{x_4}
x_5	$[0.0, 100.0]$	mm/d	H_2	SoilParameterList	MAX_BASEFLOW_RATE TOPSOIL
x_6	$[0.5, 2.0]$	-	H_1, H_2	SoilParameterList	BASEFLOW_N TOPSOIL
x_{38}	$[0.0, 1.0]$	-	H_2	SoilParameterList	BASEFLOW_THRESH TOPSOIL
Soil evaporation (AET):					
x_8	$[0.0, 3.0]$	-	$I_1 - I_3$, Q_1	SoilParameterList	PET_CORRECTION TOPSOIL
x_9	$[0.0, 0.05]$	frac	I_2,I_3	SoilParameterList	SAT_WILT TOPSOIL
x_{10}	$[0.0, 0.45]$	frac	I_2,I_3	SoilParameterList	FIELD_CAPACITY TOPSOIL = SAT_WILT TOPSOIL + x_{10}
Baseflow:					
x_{11}	$[-5.0, -2.0]$	1/d	P_1	SoilParameterList	BASEFLOW_COEFF PHREATIC = $10.0^{x_{11}}$
x_{12}	$[0.5, 2.0]$	-	P_1, P_2	SoilParameterList	BASEFLOW_N PHREATIC
x_{39}	$[0.0, 1.0]$	-	P_2	SoilParameterList	BASEFLOW_THRESH PHREATIC
x_{44}	$[0.5, 100.0]$	mm/d	P_2	SoilParameterList	MAX_BASEFLOW_RATE PHREATIC
Potential melt:					
x_{24}	$[1.5, 3.0]$	mm/d/°C	L_1	LandUseParameterList	MIN_MELT_FACTOR
x_{25}	$[0.0, 5.0]$	mm/d/°C	L_1	LandUseParameterList	MAX_MELT_FACTOR = MIN_MELT_FACTOR + x_{25}
x_{25}	$[0.0, 5.0]$	mm/d/°C	L_2	LandUseParameterList	MELT_FACTOR = MIN_MELT_FACTOR + 0.5*x_{25}

Continued on next page

Table 7.6 – *Continued from previous page*

Param.	Range	Unit	Proc. Opt.	Raven table	Parameter name
x_{26}	$[-1.0, 1.0]$	°C	L_1,L_2	LandUseParameterList	DD_MELT_TEMP
x_{27}	$[0.01, 0.2]$	1/mm	L_1	LandUseParameterList	DD_AGGRADATION
Infiltration:					
x_1	$[0.0, 1.0]$	-	M_1	LandUseParameterList	HMETS_RUNOFF_COEFF
Snow balance:					
x_{18}	$[0.0, 5.0]$	mm/d/°C	N_1	LandUseParameterList	REFREEZE_FACTOR
x_{19}	$[0.0, 0.4]$	frac	N_1	GlobalParameter	SNOW_SWI
Canopy interception:					
x_{49}	$[0.0, 0.2]$	-	O_1	VegetationParameterList	RAIN_ICEPT_PCT
x_{56}	$[1.0, 1.5]$	-	O_1	VegetationParameterList	SNOW_ICEPT_PCT $= x_{49} * x_{56}$
Canopy drip:					
x_{52}	$[0.0, 10.0]$	-	O_1,P_1,Q_1	VegetationParameterList	MAX_CAPACITY
x_{56}	$[1.0, 1.5]$	-	O_1,P_1	VegetationParameterList	MAX_SNOW_CAPACITY $= x_{52} * x_{56}$
Canopy evaporation:					
x_{48}	$[0.0, 0.99]$	-	Q_1	LandUseParameterList	FOREST_SPARSENESS
Abstraction:					
x_{41}	$[0.0, 1.0]$	-	R_1	LandUseParameterList	ABST_PERCENTAGE
Depression seepage:					
x_{40}	$[-1.0, -3.0]$	1/d	S_1	LandUseParameterList	DEP_SEEP_K $= 10.0^{x_{40}}$
Groundwater upwelling:					
x_{53}	$[0.0, 6.0]$	mm/d	T_1	SoilParameterList	MAX_CAP_RISE_RATE TOPSOIL
x_{54}	$[0.0, 12.0]$	mm/d	T_1	SoilParameterList	MAX_CAP_RISE_RATE PHREATIC
Percolation:					
x_{28}	$[0.00001, 0.02]$	1/d	U_1	SoilParameterList	PERC_COEFF TOPSOIL
x_{35}	$[0.01, 5.0]$	-	U_1	SoilParameterList	PERC_COEFF PHREATIC
Convolution (surface runoff):					
x_{20}	$[0.3, 20.0]$	-	V_1	LandUseParameterList	GAMMA_SHAPE
x_{21}	$[0.01, 5.0]$	-	V_1	LandUseParameterList	GAMMA_SCALE
Convolution (delayed runoff):					
x_{22}	$[0.5, 13.0]$	-	W_1	LandUseParameterList	GAMMA_SHAPE2
x_{23}	$[0.15, 1.5]$	-	W_1	LandUseParameterList	GAMMA_SCALE2
Rain-snow partitioning:					
x_{31}	$[-3.0, 3.0]$	°C	X_1	GlobalParameter	RAINSNOW_TEMP
x_{32}	$[0.5, 4.0]$	°C	X_1	GlobalParameter	RAINSNOW_DELTA
Precipitation correction:					
x_{33}	$[0.8, 1.2]$	-	Y_1	Gauge	RAINCORRECTION
x_{34}	$[0.8, 1.2]$	-	Y_1	Gauge	SNOWCORRECTION
Soil model:					

Continued on next page

Table 7.6 – *Continued from previous page*

Param.	**Range**	**Unit**	**Proc. Opt.**	**Raven table**	**Parameter name**
x_{29}	$[0.0, 0.5]$	m	$H_{1,2}$,$I_{1,2,3}$ M_1,U_1	SoilProfiles	thickness TOPSOIL L
x_{30}	$[0.0, 2.0]$	m	$J_{1,2}$, U_1	SoilProfiles	thickness PHREATIC
LAI model:					
x_{55}	$[0.0, 1.0]$		K_3,O_1, P_1,Q_1	SeasonalRelativeLAI	LAI seasonal decrease factor

Table 7.7 Values from individual regressions applied to data separately within each MOPEX catchment correlating validation KGE to the number of parameters across multiple independent trials and model configurations. All p-values less than the 10% significance level or shown in **bold**. Regression was a basic linear regression of the form $Y \sim \beta_1 X + \beta_0$ and the p-values are associated with null hypothesis of slope $\beta_1 = 0$.

Index	Catchment ID (State)	slope (β_1)	intercept (β_0)	slope p-value
1	01608500 (West Virginia)	-9.83e-04	0.651	0.148
2	01643000 (Maryland)	3.88e-04	0.801	0.570
3	01668000 (Virginia)	1.12e-03	0.817	**0.066**
4	03054500 (West Virginia)	2.61e-04	0.858	0.317
5	03179000 (West Virginia)	-5.17e-04	0.883	0.545
6	03364000 (Indiana)	-1.17e-03	0.884	0.188
7	03451500 (North Carolina)	1.1e-04	0.828	0.886
8	05455500 (Idaho)	-7.37e-05	0.596	0.964
9	07186000 (Missouri)	8.92e-04	0.827	0.144
10	07378500 (Louisiana)	8.46e-04	0.859	0.206
11	08167500 (Texas)	1.14e-03	0.533	0.375
12	08172000 (Texas)	1.78e-03	0.667	0.171

Table 7.8 Values from individual regressions applied to data separately within each MOPEX catchment correlating the validation gap (maximum calibration KGE minus validation KGE of the same model run) to the number of parameters across multiple independent trials and model configurations. All p-values less than the 10% significance level are shown in **bold**. Regression was a basic linear regression of the form $Y \sim \beta_1 X + \beta_0$ and the p-values are associated with null hypothesis of slope $\beta_1 = 0$.

Index	Catchment ID (State)	slope (β_1)	intercept (β_0)	slope p-value
1	01608500 (West Virginia)	1.12e-03	0.232	0.111
2	01643000 (Maryland)	-4.69e-05	0.104	0.938
3	01668000 (Virginia)	-8.73e-04	0.071	**0.066**
4	03054500 (West Virginia)	-1.33e-05	0.027	0.966
5	03179000 (West Virginia)	5.54e-04	0.023	0.449
6	03364000 (Indiana)	1.23e-03	0.032	**0.081**
7	03451500 (North Carolina)	2.67e-04	0.099	0.666
8	05455500 (Idaho)	4.56e-04	0.268	0.718
9	07186000 (Missouri)	2.2e-05	0.049	0.975
10	07378500 (Louisiana)	-3.28e-04	0.046	0.450
11	08167500 (Texas)	-5.23e-04	0.303	0.712
12	08172000 (Texas)	-1.34e-03	0.129	0.250

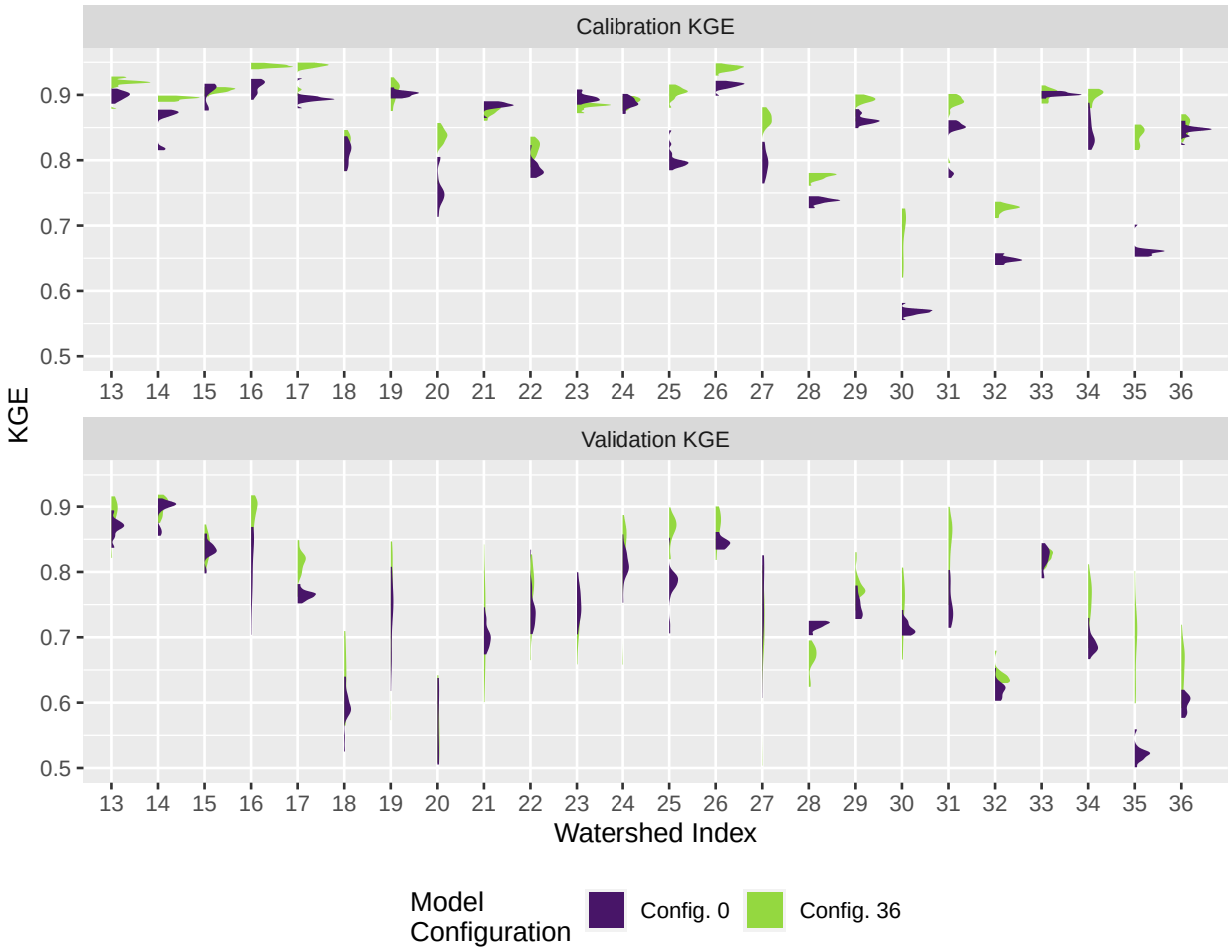


Figure 7.4 Calibration performance distributions, generated from 20 independent trials of the calibration for model configurations 0 (original) and 36 (version 2) for 24 independent catchments.

Table 7.9 Results of the unpaired, two-sided Wilcoxon rank sum test in the independent validation catchments shown. A confidence level of 95% is used. The median KGE performance for both the original blended model configuration 0 and the selected blended model configuration 36 are shown, as well as the p-values for testing whether model config. 0 is different than the model config. 36.

Watershed Index	Model Config. 0 Median KGE	Model Config. 36 Median KGE	p-value
Calibration KGE comparisons			
13	0.900	0.919	3.94e-08
14	0.871	0.895	6.79e-08
15	0.912	0.908	1.37e-01
16	0.917	0.944	1.45e-11
17	0.894	0.946	1.02e-10
18	0.816	0.824	1.27e-01
19	0.902	0.911	1.48e-01
20	0.751	0.838	2.76e-10
21	0.885	0.878	1.19e-04
22	0.783	0.820	2.32e-08
23	0.894	0.884	2.00e-06
24	0.887	0.891	4.95e-01
25	0.797	0.906	1.45e-11
26	0.916	0.941	1.45e-11
27	0.796	0.864	1.02e-10
28	0.738	0.777	1.45e-11
29	0.861	0.892	1.76e-08
30	0.568	0.705	1.45e-11
31	0.850	0.890	1.33e-08
32	0.647	0.728	1.45e-11
33	0.901	0.903	6.40e-01
34	0.833	0.901	1.02e-10
35	0.660	0.841	1.45e-11
36	0.847	0.852	7.18e-02
Validation KGE comparisons			
13	0.871	0.894	7.47e-04
14	0.902	0.905	2.21e-01
15	0.834	0.837	4.45e-01
16	0.836	0.898	2.83e-09
17	0.765	0.818	5.41e-09
18	0.592	0.639	1.12e-02
19	0.750	0.771	2.63e-02
20	0.494	0.405	4.91e-02
21	0.701	0.710	7.58e-01
22	0.739	0.781	4.60e-02
23	0.749	0.738	2.65e-01
24	0.810	0.848	2.39e-03
25	0.787	0.870	1.74e-10
26	0.845	0.877	3.36e-05
27	0.761	0.724	2.11e-02
28	0.720	0.672	1.45e-11
29	0.744	0.772	3.36e-05
30	0.711	0.756	1.59e-03
31	0.743	0.857	5.41e-09
32	0.621	0.638	3.03e-08
33	0.826	0.824	8.83e-01
34	0.686	0.761	2.12e-07
35	0.515	0.704	1.45e-11
36	0.601	0.658	1.33e-08

Table 7.10 Mean calibration and validation KGE performances for the original and v2 blended model versions for each of the 36 catchments utilized in this study. The differences in mean calibration and validation performances are also calculated.

Basin ID	State	Original Calib KGE	v2 Calib KGE	Calib Diff	Original Valid KGE	v2 Valid KGE	Valid Diff
1	WV	0.815	0.885	0.07	0.530	0.589	0.0589
2	MD	0.893	0.916	0.0231	0.775	0.833	0.0579
3	VA	0.834	0.894	0.0604	0.769	0.868	0.0988
4	WV	0.856	0.899	0.0431	0.854	0.866	0.0115
5	WV	0.830	0.911	0.081	0.769	0.869	0.0997
6	IN	0.866	0.920	0.0544	0.779	0.844	0.065
7	NC	0.907	0.946	0.0385	0.776	0.859	0.0829
8	IA	0.792	0.883	0.0914	0.369	0.608	0.239
9	MO	0.789	0.916	0.126	0.864	0.897	0.0326
10	LA	0.896	0.927	0.0307	0.853	0.905	0.0517
11	TX	0.796	0.848	0.0517	0.650	0.649	-1.11e-03
12	TX	0.762	0.812	0.0497	0.742	0.798	0.0566
13	ME	0.899	0.918	0.0192	0.868	0.888	0.0203
14	CT	0.861	0.895	0.0332	0.895	0.901	6.43e-03
15	NY	0.904	0.907	3.52e-03	0.834	0.838	4.35e-03
16	MS	0.914	0.944	0.0309	0.820	0.892	0.0715
17	TN	0.895	0.942	0.0475	0.765	0.814	0.0492
18	UT	0.812	0.820	8e-03	0.595	0.635	0.0403
19	CA	0.901	0.905	3.36e-03	0.736	0.768	0.0318
20	CA	0.756	0.834	0.0779	0.514	0.300	-0.214
21	CA	0.884	0.877	-7.31e-03	0.703	0.713	0.0103
22	ID	0.787	0.818	0.0312	0.743	0.756	0.0129
23	ID	0.895	0.884	-0.0106	0.753	0.738	-0.0142
24	ID	0.888	0.890	1.49e-03	0.811	0.834	0.0222
25	VA	0.800	0.904	0.104	0.782	0.866	0.0842
26	GA	0.915	0.941	0.026	0.845	0.871	0.026
27	FL	0.796	0.862	0.0655	0.759	0.718	-0.0417
28	KY	0.738	0.775	0.0376	0.718	0.669	-0.0492
29	OH	0.862	0.890	0.0277	0.747	0.777	0.0299
30	WI	0.568	0.688	0.12	0.698	0.746	0.0471
31	IL	0.838	0.886	0.0482	0.751	0.845	0.0941
32	KS	0.648	0.727	0.0784	0.619	0.639	0.0206
33	ID	0.901	0.901	-5.97e-04	0.821	0.824	2.35e-03
34	AR	0.843	0.899	0.0556	0.690	0.754	0.0643
35	KS	0.661	0.838	0.177	0.514	0.691	0.177
36	OR	0.846	0.852	5.93e-03	0.599	0.656	0.0573