https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

Advancement of a blended hydrologic model for robust model performance

Robert Chlumsky[1], Juliane Mai[2,3], James R. Craig[1], and Bryan A. Tolson[1]

[1]Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, ON, Canada
[2]Computational Hydrosystems, Helmholtz Centre for Environmental Research - UFZ, Leipzig, Saxony, Germany
[3]Center for Scalable Data Analytics and Artificial Intelligence - ScaDS.AI, Leipzig, Saxony, Germany

Correspondence: Robert Chlumsky (robert.chlumsky@uwaterloo.ca)

Abstract. A blended model structure has emerged as an alternative to the traditional representation of model structure in a hydrologic model, in which multiple algorithmic choices are used to represent some hydrologic process within a model, and are combined within a single model run using a weighted average of process fluxes. This approach has been shown to improve overall model performance, as well as provide an efficient way to test multiple model structures. We propose that a blended model may also be at least a partial solution to the calls for a more robust Community Hydrologic Model, which can mitigate the need for developing new hydrologic models for each catchment and application.

We develop an updated version of the blended model configuration which defines the suite of all possible hydrologic process options in the blended model. Configuration development was guided by model performance for more than 30 different discrete model configurations across 12 MOPEX catchments. Improvements to the blended model include the introduction of blended potential melt and potential evapotranspiration as new process groups, inclusion of non-blended structural changes, and a revision of the process options within each existing group. This leads to a very high-performing model with a mean calibration Kling-Gupta Efficiency (KGE) score of 0.90 and mean validation KGE score of 0.80 across all 12 MOPEX catchments, a substantial improvement in model performance relative to the initial version of 0.06 and 0.07 in calibration and validation, respectively. We test for overfitting of models and find little statistical evidence that increasing the complexity of blended models reduces validation performance. We then select the preferred model configuration as version 2 of the blended model, and test it with 12 independent catchments, which shows a mean calibration and validation score of 0.89 and 0.76, respectively, and improvement over the original model (0.03 in mean calibration KGE score). Version 2 of the blended model is robust across a range of catchments without the need for adjusting its flexible model structure, and may be useful in future hydrology studies and applications alike.

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

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023


Hydrology and Earth System Sciences
Discussions

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.

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., 2022c), 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., 2022a), 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., 2022b; 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., 2022c) further encourages hydrologists pushing for process-based approaches to confront these deficiencies in their models.

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

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023

Hydrology and Earth System Sciences
Discussions

dominant deep-learning model (Mai et al., 2022c). These results have been further confirmed when more than 600 experts where tasked to rate these hydrographs without revealing the model that was used to produce the hydrograph (Gauch et al., 2022).

In this study, we improve 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. (2022c) with the goal of producing an improved blended model. This goal is distinct from building an optimal model for a single application and catchment; ~~this~~ intention is to design a blended structure that can be robust across many applications and catchments. 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 of the manuscript are organized as follows. The methods (Section 2) is organized into several sections. Section 2.1 introduces the concept of blended models in general and provides the theory necessary to understand the subsequent model developments. Section 2.2 discusses the update to the initial blended model configuration, including a description of the developments and the empirical approach taken. Section 2.3 describes the tests used to check for model overfitting. Section 2.4 describes the metrics used to aid in the selection of a single preferred blended model structure from the multiple candidate configurations. Section 2.5 describes the methodology for validation of the selected blended model version 2 against the original version in an independent set of catchments. Results and Discussion (Section 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.

2 Methods

The following methodology sections are organized as follows. **Section 2.1** introduces the blended modelling approach and provides the necessary background on the original blended model. **Section 2.2** discusses the iterative development and evaluation of candidate blended model configurations. **Section 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 2.4** discusses the selection process for a new version of the blended model from the set of plausible blended model configurations. Finally, **Section 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.

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

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023


Hydrology and Earth System Sciences
Discussions
Open Access
EGU

included five process 'groups': infiltration, quickflow, 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:

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

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 $\mathcal{G}$.

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.

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., 2022a) 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.

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. The continuous weighting of process options in a blended model allows model structure to be expressed as a continuum of plausible options, rather than discrete points represented by separate model structures. If we imagine that optimal model solutions may exist within this continuum, rather than precisely at a specific model structure defined by discrete options, 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. (2022c).

2.2 Blended model development and performance evaluation

This study deploys a blended model in over 30 model configurations, with up to seven blended process groups used in any given model configuration. In the model development phase, blended model configurations are tested in calibration mode using

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023


Hydrology and Earth System Sciences
Discussions

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.

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. The Kling-Gupta Efficiency (KGE) metric (Gupta et al., 2009) is used as the objective function. 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).

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 successive each 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 (Chlumsky et al., 2021, such as Fig. 5 therein), 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.

The model development process involved (a) the adjustment of selected process equations (Section 2.2.1), (b) addition of new blended process groups (Section 2.2.2), (c) testing of the so-called conglomerate model, (d) changes to the non-blended part of the model structure (Section 2.2.4), and (d) the reduction of complexity by removing process equations with marginal contribution to performance (Section 2.2.5). These specific developments are discussed in the following subsections. The specific algorithms included in process groups for each blended model configuration is provided in **Table A1** of the Appendix.

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023


Hydrology and Earth System Sciences
Discussions

2.2.1 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, recognizing that the original blended configuration selected some options that worked well based on variations from the HMETS model (Martel et al., 2017) rather than options that would maximize robustness and coverage of model space. In cases where process options of a similar form or function exist 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.

2.2.2 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 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.

2.2.3 Conglomerate model

A so-called conglomerate model (model configuration 24) is tested and calibrated in the 12 MOPEX catchments. 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 indicate whether the maximization of the number of process options included in the blended model configuration is advantageous relative to more carefully curated model configurations.

2.2.4 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

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023


Hydrology and Earth System Sciences
Discussions

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.

2.2.5 Reduction in complexity

Attempts are made between the model configuration iterations to reduce the complexity 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 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 either calibration or validation across the 12 tested MOPEX catchments is generally regarded as a benefit. This strategy of reducing model complexity was employed throughout the model development stages where possible to maintain parsimony.

2.3 Model overfitting

One of the assumptions made in this study is that increasing the complexity and the number of model parameters 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, and an increasing inability to identify optimal model solutions.

In order to 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 catchments, and repeated within each of the 12 MOPEX catchments. 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 (defined as the maximum calibration KGE score minus the validation KGE score of the same trial) 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.

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 evaluate relative performance in model configurations, minute discrepancies in this metric may not be the most informative when

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023


Hydrology and Earth System Sciences
Discussions

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 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 metrics, 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.

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 12 additional catchments, 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 12 independent catchments were selected randomly from the set of catchments calibrated by Mai et al. (2022b), with the criteria that they were also 1) 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. The map of the 12 MOPEX and the 12 independent catchments is provided in **Figure 1**. Additional information on the selected catchments is provided in **Table 1**.

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

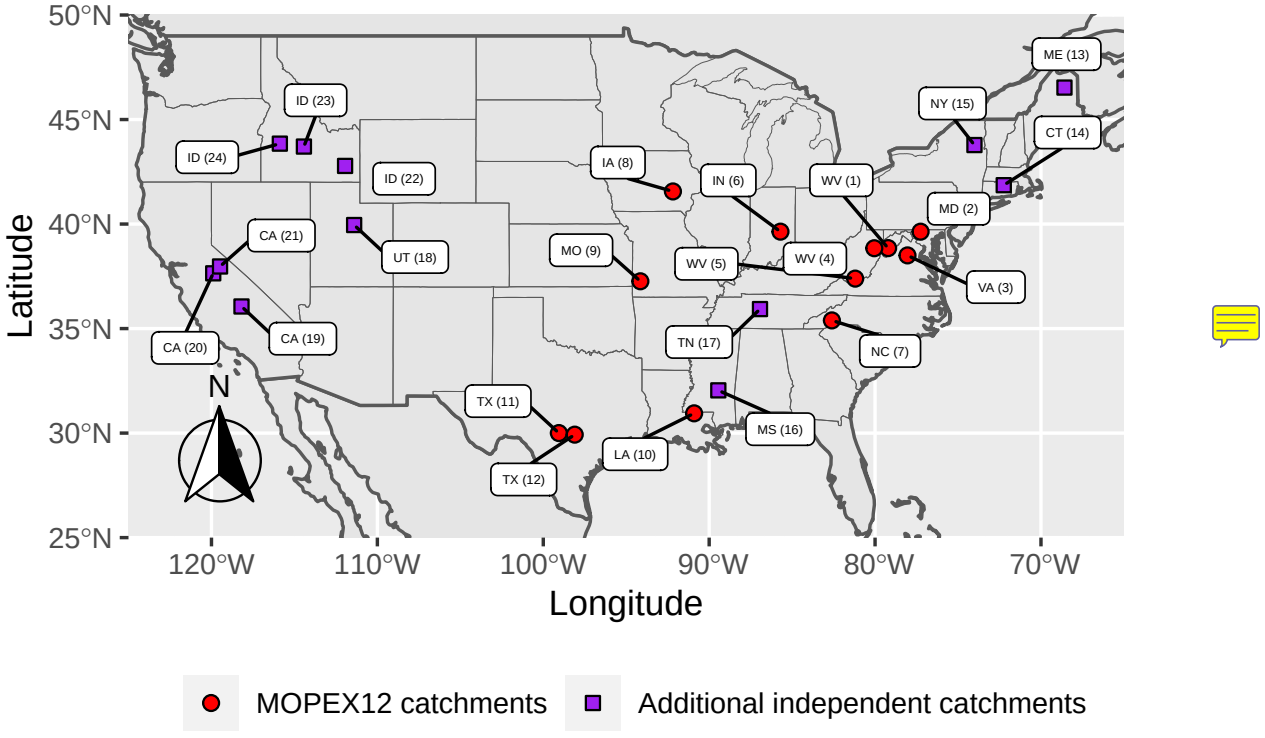

Figure 1. Map of the 12 MOPEX12 catchments located in the southeastern part of the United States, as well as 12 additional independent catchments selected to further test the blended model version 2 (model configuration 36). Captions indicate the state abbreviation; the number in parentheses indicates the catchment index (i.e., 1 to 24).

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023

Hydrology and Earth System Sciences
Discussions

Table 1. Independent catchments from HYSETS selected for validating the blended model version 2

Index	Catchment ID	Catchment Name	Area (km^2)
13	01017000	AROOSTOOK RIVER AT WASHBURN, ME	4282.0
14	01122500	SHETUCKET RIVER NEAR WILLIMANTIC, CT	1045.9
15	01318500	HUDSON RIVER AT HADLEY NY	4307.9
16	02472000	LEAF RIVER NR COLLINS, MS	1923.5
17	03434500	HARPETH RIVER NEAR KINGSTON SPRINGS, TN	1768.2
18	10150500	SPANISH FORK AT CASTILLA, UT	1688.0
19	11189500	SF KERN R NR ONYX CA	1372.1
20	11272500	MERCED R NR STEVINSON CA	3295.6
21	11276500	TUOLUMNE R NR HETCH HETCHY CA	1183.1
22	13073000	PORTNEUF RIVER AT TOPAZ ID	1491.2
23	13139500	BIG WOOD RIVER AT HAILEY ID	1656.9
24	13200000	MORES CREEK AB ROBIE CREEK NR ARROWROCK DAM ID	1027.8

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

3 Results and Discussion

The following results and discussion sections are organized as follows. **Section 3.1** presents a summary of the developments in each blended model configuration and initial performance results. **Section 3.2** presents the statistical analysis undertaken assess model overfitting. **Section 3.3** presents the selection process to determine version 2 of the blended model from the set of blended model configurations. **Section 3.4** describes the process to validate the blended model version 2 with a set of additional independent catchments. Finally, **Section 3.5** provide additional discussion on how blended models can be used to assess structural identifiability for scientific questions.

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 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 A1** of the

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023


Hydrology and Earth System Sciences
Discussions

Appendix. Further descriptions of the process options in Raven may found in the Raven User's Manual (Raven Development Team, 2022).

1. Model configuration 0: original model version from Mai et al. (2020);
2. Model configuration 1 - original model version but using data sourced from MOPEX/HYSETS rather than PET_OUDIN;
3. Model configuration 2: initial update of blended process options from the original blended model version based on expert consideration;
4. Model configuration 3-5: addition of depression storage & seepage, followed by adjustments to other process options;
5. Model configuration 6-23: introduction of blended forcings, 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;
8. Model configuration 34-37: addition of non-blended processes (including canopy processes), and additional complexity reduction experiments.

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 possible 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 2**. The distribution of performance results for select model configurations is provided in **Figure 3**, which also shows a general improvement in performance with model configuration.

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

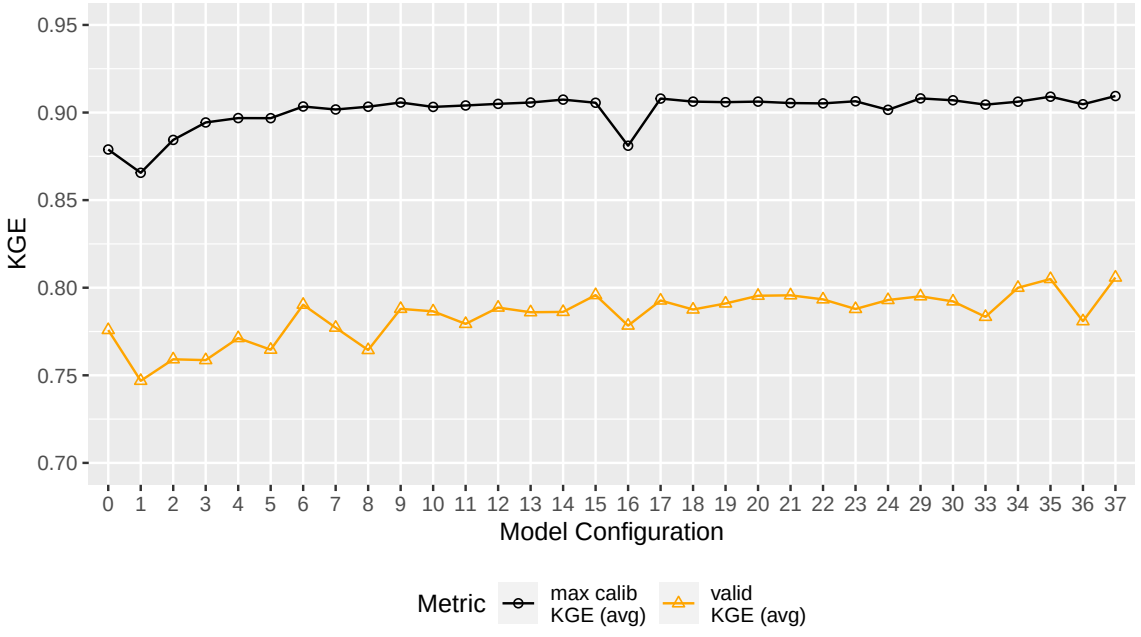

Figure 2. Model performance in calibration and validation by model configuration. The maximum calibration KGE is the maximum KGE within a given catchment across 20 independent trials. The validation KGE is the KGE associated with the maximum calibration KGE score, where the model setup with the maximum calibration KGE score is used to perform a temporal validation simulation. The values shown in this figure are the maximimum calibration KGE and validation KGE averaged across 12 MOPEX catchments, and plotted for each blended model configuration.

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023

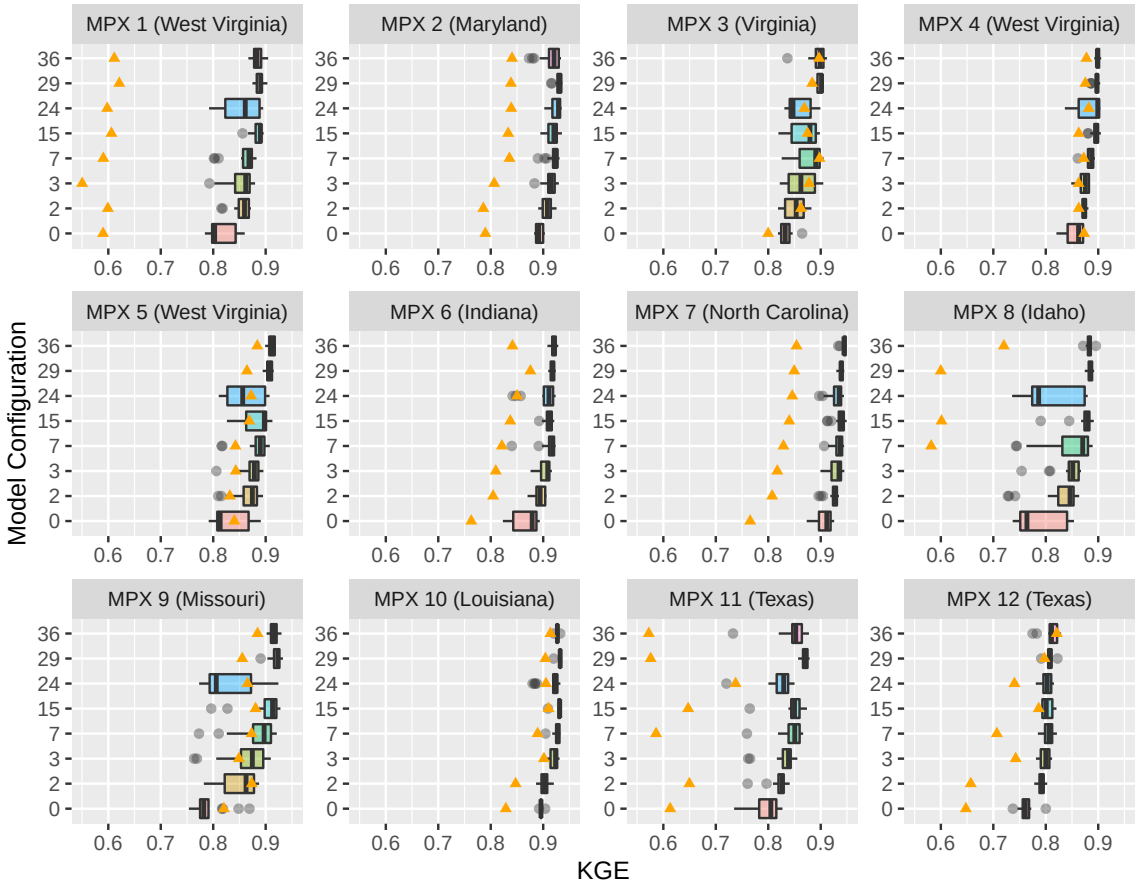

Figure 3. Calibration performance distributions, generated from 20 independent trials of the calibration for each model configuration and MOPEX catchment, shown for select model configurations. The orange points indicate the validation performance of trial with the maximum calibration performance within the 20 independent trials.

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 catchments, and second to individual catchments. The overall regression for all model configurations and catchments is shown in **Figure 4**, and the individual regressions are summarized in **Table 2**.

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

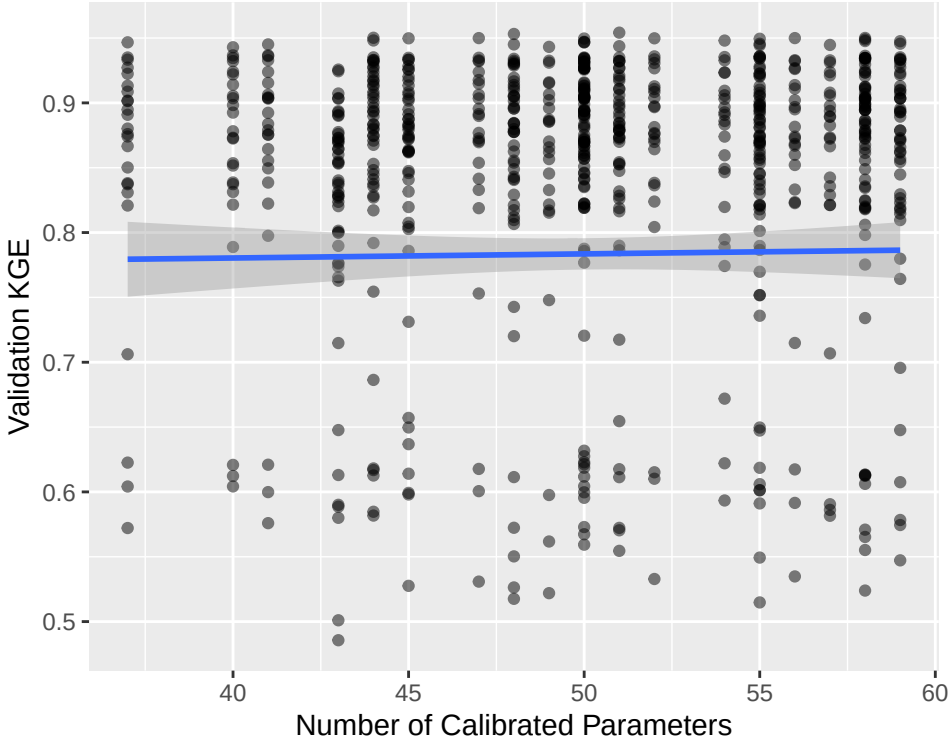

Figure 4. Validation KGE on the y-axis plotted against the number of calibrated model parameters on the x-axis. Each of the 372 data points represents the validation score associated with the top-performing calibrated model of 20 independent trials, for each of 31 model configurations 12 MOPEX catchments ($31 \times 12 = 372$). 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.

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023

Hydrology and Earth System Sciences
Discussions

Table 2. 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

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023


Hydrology and Earth System Sciences
Discussions

Figure 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 applied 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 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 3**.

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023

Hydrology and Earth System Sciences
Discussions

Table 3. 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

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023


Hydrology and Earth System Sciences
Discussions

Table 3 shows 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.

3.3 Selection of a preferred blended model configuration

The sequential improvement of the overall model performance is shown in **Figure 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 2.4) which are plotted in **Figure 5**. In both plots A and B, the ideal point is set in the bottom left corner of the plot.

For **Figure 5A**, this ideal point is located at $[0,0]$, and represents the point where 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 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 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 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 noteworthy 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).

In **Figure 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 noteworthy 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 selection is dependent on the setup of this study, such as the daily timestep, lumped model discretization,

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

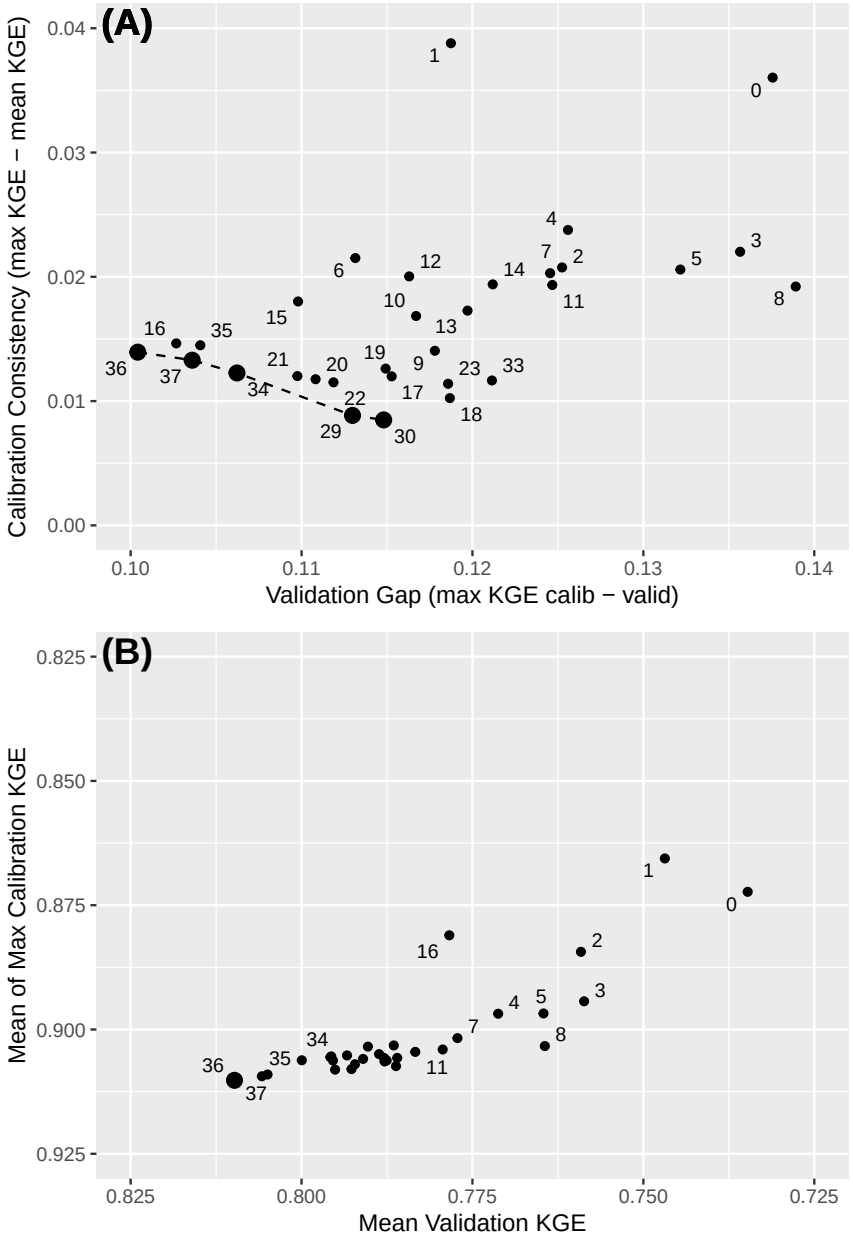

Figure 5. A) Model configurations plotted based on the consistency in model calibration performance on the y-axis (mean difference in the maximum calibration KGE minus mean calibration KGE for 20 independent trials, averaged across all 12 MOPEX catchments) against the 'validation gap' on the x-axis, measured as the maximum KGE calibration performance (taken as the best solution of 20 independent trials) minus its associated validation performance, averaged across 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) Model configurations plotted based on the mean of the maximum calibration KGE performance, averaged across all 12 catchments) against the mean validation KGE performance for the model iteration with the maximum calibration KGE on the x-axis, averaged across catchments. Model configuration 36 is the only non-dominated model configuration in this plot. In both plots, Pareto points are plotted with larger shapes than other data points.

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023


Hydrology and Earth System Sciences
Discussions

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 provided in **Table A2**, and the parameter information for model configuration 36 is provided in **Table A3** of the Appendix. Raven model files are also available with the supporting data for model configuration 36 (see the Code and Data Availability section).

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 12 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 using a combination of density and box plots for calibration performance, and the validation period performance associated with the best calibration solution, in **Figure 6**.

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023

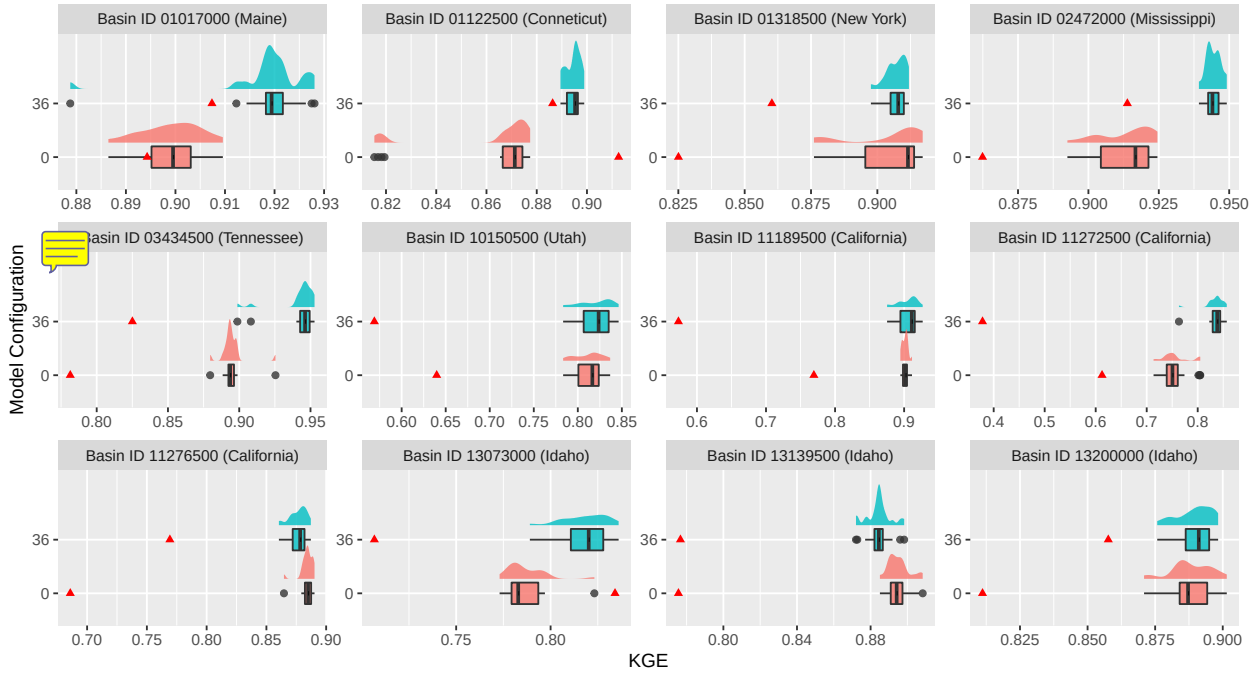

Figure 6. Calibration performance distributions and boxplots, generated from 20 independent trials of the calibration for model configurations 0 (original) and 36 (version 2) for 12 independent catchments. The orange point indicates the validation performance of the best calibration solution from the 20 independent trials.

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023


Hydrology and Earth System Sciences
Discussions

From the results, model configuration 36 improved the calibration performance in 8 of 12 catchments compared to model configuration 0, with a similar performance in an additional 2 catchments. The overall mean calibration KGE improved from approximately 0.866 to 0.886 moving from the original blended model configuration to model configuration 36, and improved just 0.003 regarding KGE in mean validation, effectively signaling no change in validation performance for the additional catchments. In considering the mean of the best run in each catchment, the KGE improved from 0.886 to 0.899, and the validation KGE improved from 0.803 to 0.834.

The initial high performance of the original blended model version is noteworthy, with all calibrated model performances with the original blended model (configuration 0) exceeding a KGE of 0.7, and falling between 0.85 and 0.90 KGE in most catchments. The calibration and (temporal) validation using a substantially reduced calibration budget of 2000 model evaluations (instead of 10,000) and 10 independent trials (instead of 20) across more than 3000 catchments in North America lead to a median NSE performance of 0.73 and 0.64, respectively (Mai et al., 2022b, see Fig. 2 therein).

Overall, the work presented here indicates an improvement in the model calibration performance with blended model configuration 36 over the original model configuration 0, but suggests that the increase in validation performance may be less than that seen in the MOPEX 12 catchments where the new blended model version was developed.

3.5 Identifiability in blended models

In the development of the new blended model configuration and testing multiple process algorithms, distributions of weights for each process option were continuously examined to determine which options were being selected for. In borrowing the concept of 'identifiability' from literature in being able to select a unique parameter value from the data available (Wagener et al., 2003; Guillaume et al., 2019), and applying it to model structure, we can informally asses the identifiability of a process group in a blended model as more or less identifiable based on the broadness of weight distributions for a given model configuration. This was demonstrated in Chlumsky et al. (2021) in examining box plots of model weights with the blended model. In this study, the distributions for the baseflow group tend to be quite wide and thus it was difficult to 'identify' a preferred option between the two baseflow algorithms, while the distributions for groups such as infiltration and PET tended to be much narrower and show some selection across trials. An interesting note is that while there was little identifiability in the baseflow group, reducing the low identifiability groups to a single option still reduced the performance of the model. This suggests that high identifiability in process groups is not required for the model to benefit in model performance from the blending of those groups - flexibility may be more important than identification in this case. This may be related to the mixing of faster and slower subsurface flow signatures, where a blended group can generate a much more complex signature than a single process.

The identifiability of process groups is very likely impacted by the choice of optimization objective, highlighting the difficulty in selecting a single optimization metric. In this study the KGE was used, which likely allows some process groups, such as infiltration or snowbalance, to be quite identifiable, and likely obscures the identifiability of processes that respond on a different timescale, such as baseflow. A low flow metric would likely improve the identifiability of the baseflow group, though this remains to be tested.

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023


Hydrology and Earth System Sciences
Discussions

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 potential 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 further evaluated against an additional 12 independent catchments within the continental United States, and generally found to improve model performance relative to the original blended model configuration in most of the catchments even though validation performances are not as pronounced as was seen using the 12 MOPEX catchments used to develop the revised model. 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, assessing the identifiability of these processes, new options for blended forcings, and strategies to reduce the dimensionality 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. 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 identifiability of processes in hydrologic models.

Code and data availability. The code and data used for this analysis will be available on Github (https://github.com/rchlumsk/blendedmodel_update_2022) upon publication of this manuscript. The Raven Hydrologic Modelling Framework v3.6 is available at http://raven.uwaterloo.ca/Downloads.html. The DDS algorithm and Ostrich software v21.03.16 are available at https://github.com/usbr/ostrich/releases/tag/v21.03.16.

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023


Hydrology and Earth System Sciences
Discussions

Appendix A: Additional information on the blended model configurations

This appendix contains a table showing the process options included within each blended model configuration tested (**Table A1**), a table listing all of the processes and process options included in the selected blended model version 2 (**Table A2**), and a table listing all model parameters included in the selected blended model version 2 (**Table A3**).

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

Table A1. Process options included within each blended model configuration. Model configurations 25-28 & 31-32 were used for testing weighting schemes and are omitted.

	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 from Mai et al. (2020)																																					
0	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				43
3. Initial update based on expert consideration of process options																																					
2	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	48
4	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				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	X			X	57
8	X	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		X		58
10	X	X		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	X		X		X		56
12	X	X		X		X	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	X		X		X		58
14	X	X		X		X	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	X	X		X		X		55
16	X	X		X		X	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		X		52
18	X		X	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	X	X		50
20	X	X		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	X		50
22	X	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		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	79
7. Experiments in reducing complexity of the blended configuration																																					
29	X		X	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	X				X		40
33	X		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				X				X		45
35	X		X	X			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	X				X			48
37	X		X	X				X					X				X	X			X			X				X				X				44

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023

Hydrology and Earth System Sciences
Discussions

Table A2. 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 A3. 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	–

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023

Hydrology and Earth System Sciences
Discussions

Table A3: 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}
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

Continued on next page

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023

Hydrology and Earth System Sciences
Discussions
Open Access
EGU

Table A3 – *Continued from previous page*

Param.	Range	Unit	Proc. Opt.	Raven table	Parameter name
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:					
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

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023


Hydrology and Earth System Sciences
Discussions

Author contributions. RC set up the analyses, implemented blended forcing options in Raven, iterated through the model configurations, wrote the majority of the manuscript, and prepared all figures and tables. JM contributed to the writing of the manuscript, provided initial blended model setup files and data including the HYSETS data, and helped with the interpretation of results. JRC contributed to the writing of the manuscript, helped with the analysis and interpretation of results, suggested non-blended structural model adjustments, reviewed the model configuration for version 2, and verified the code changes in Raven. BAT contributed to the interpretation of results, the writing of the manuscript, and the verification of appropriate parameter ranges in blended model configurations. All co-authors contributed to the initial expert selection of alternative process options for blended model configuration 2.

Competing interests. The authors declare that they have no conflict of interest

Acknowledgements. This research has been supported by the Natural Sciences and Engineering Research Council of Canada (grant no. CGSD3-558879-2021) and the Engineering Excellence Doctoral Fellowship provided at the University of Waterloo. Dr. Craig and Dr. Tolson both acknowledge partial support through their NSERC Discovery Individual grants. This research was undertaken thanks in part to funding from the CANARIE research software funding program (project RS-332). The work was made possible by the facilities of the Shared Hierarchical Academic Research Computing Network (SHARCNET; www.sharcnet.ca) and Compute/Calcul Canada.

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023

References

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

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

Chlumsky, R., Mai, J., Craig, J. R., and Tolson, B. A.: Simultaneous Calibration of Hydrologic Model Structure and Parameters Using a Blended Model, Water Resources Research, 57, e2020WR029 229, https://doi.org/10.1029/2020WR029229, 2021.

Clark, M. P., Kavetski, D., and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modeling, Water Resources Research, 47, https://doi.org/10.1029/2010WR009827, 2011.

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

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

Frame, J., Kratzert, F., Gupta, H. V., Ullrich, P., and Nearing, G. S.: On Strictly Enforced Mass Conservation Constraints for Modeling the Rainfall-Runoff Process, Hydrological Processes, in review, https://doi.org/10.31223/X5BH0P, 2022a.

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

Gauch, M., Kratzert, F., Gilon, O., Gupta, H., Mai, J., Nearing, G., Tolson, B., Hochreiter, S., and Klotz, D.: In Defense of Metrics: Metrics Sufficiently Encode Typical Human Preferences Regarding Hydrological Model Performance, EarthArXiv, https://doi.org/10.31223/X52938, 2022.

Guillaume, J. H., Jakeman, J. D., Marsili-Libelli, S., Asher, M., Brunner, P., Croke, B., Hill, M. C., Jakeman, A. J., Keesman, K. J., Razavi, S., and Stigter, J. D.: 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, 119, 418–432, 2019.

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

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

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

https://doi.org/10.5194/hess-2023-69
Preprint. Discussion started: 23 March 2023

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

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

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

Mai, J., Craig, J. R., Tolson, B. A., and Arsenault, R.: The sensitivity of simulated streamflow to individual hydrologic processes across North America, Nature communications, 13, 455–455, https://doi.org/10.1038/s41467-022-28010-7, 2022b.

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

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

Matott, L. S.: 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 edn., 2017.

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

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., Prieto, C., and Gupta, H. V.: What Role Does Hydrological Science Play in the Age of Machine Learning?, Water Resources Research, 57, e2020WR028 091, https://doi.org/10.1029/2020WR028091, 2021.

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

Perrin, C., Michel, C., and Andréassian, V.: Improvement of a parsimonious model for streamflow simulation, Journal of Hydrology, 279, 275–289, https://doi.org/10.1016/S0022-1694(03)00225-7, 2003.

Raven Development Team: Raven: User's and Developer's Manual v3.5, http://raven.uwaterloo.ca/, 2022.

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

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

Weiler, M. and Beven, K.: Do we need a Community Hydrological Model?, Water Resources Research, 51, 7777–7784, https://doi.org/10.1002/2014WR016731, 2015.