**RESEARCH ARTICLE**
10.1029/2019WR027009

**Key Points:**
- Conceptual model structures can be optimized simultaneously with model parameters.
- The identified model structures are able to reproduce the rainfall runoff behavior of humid catchments.
- Standard optimization algorithms are not ideal for structure identification as sets of parameters to calibrate depend on model structure.

**Supporting Information:**
- Supporting Information S1

**Correspondence to:**
D. Spieler,
diana.spieler@tu-dresden.de

**Citation:**
Spiele, D., Mai, J., Craig, J. R., Tolson, B. A., & Schütze, N. (2020). Automatic Model Structure Identification for Conceptual Hydrologic Models. Water Resources Research, 56, e2019WR027009. https://doi.org/10.1029/2019WR027009

**Received 23 DEC 2019**  
**Accepted 29 JUL 2020**  
**Accepted article online 3 AUG 2020**

---

**Automatic Model Structure Identification for Conceptual Hydrologic Models**

Diana Spieler1, Juliane Mai2, James R. Craig2, Bryan A. Tolson2, and Niels Schütze1

1Department of Hydrosciences, TU Dresden, Dresden, Germany, 2Department of Civil and Environmental Engineering, University of Waterloo, Waterloo, Ontario, Canada

**Abstract**  
Choosing (an) adequate model structure(s) for a given purpose, catchment, and data situation is a critical task in the modeling chain. However, despite model intercomparison studies, hypothesis testing approaches with modular modeling frameworks, and continuous efforts in model development and improvement, there are still no clear guidelines for identifying a preferred model structure. By introducing a framework for Automatic Model Structure Identification (AMSI), we support the process of identifying (a) suitable model structure(s) for a given task. The proposed AMSI framework employs a combination of the modular hydrological model RAVEN and the heuristic global optimization algorithm dynamically dimensioned search (DDS). It is the first demonstration of a mixed-integer optimization algorithm applied to simultaneously optimize model structure choices (integer decision variables) and parameter values (continuous decision variables) in hydrological modeling. The AMSI framework is thus able to sift through a vast number of model structure and parameter choices for identifying the most adequate model structure(s) for representing the rainfall-runoff behavior of a catchment. We demonstrate the feasibility of the approach by reidentifying given model structures that produced a specific hydrograph and show the limits of the current setup via a real-world application of AMSI on 12 MOPEX catchments. Results show that the AMSI framework is capable of inferring feasible model structure(s) reproducing the rainfall-runoff behavior of a given catchment. However, it is a complex optimization problem to identify model structure and parameters simultaneously. The variance in the identified structures is high due to near equivalent diagnostic measures for multiple model structures, reflecting substantial model equifinality. Future work with AMSI should consider the use of hydrologic signatures, case studies with multiple types of observation data, and the use of mixed-integer multiobjective optimization algorithms.

1. Introduction

Model structure identification and parameter calibration are essential steps in the modeling chain. Both have been identified as significant sources of uncertainty (Di Baldassarre & Montanari, 2009) and triggered an abundance of papers quantifying their effect on modeling results, e.g., those cataloged in Montanari et al. (2009). However, while parameter estimation is part of most modeling workflows, model structure identification—the process of selecting an appropriate model structure for a given purpose, catchment, and data set—is not. As shown by Addor and Melsen (2019), modelers prefer to select their models based on legacy (habit, practicality, convenience) rather than adequacy (fitness for purpose). This does not mean that adequacy is not considered. It is just not tested other than by the generally laborious mechanism of model intercomparison studies. One reason for this is that testing different models (or components of models) as hypotheses, as has been proposed in order to identify a suitable model structure (Beven, 2001, 2018; Chamberlin, 1965; Clark et al., 2011), is generally a time-consuming manual process, while parameter estimation is mainly treated as an automated process.

Thus, model intercomparison studies of the last decade typically compared a rather limited amount (≤20) of fixed model structures in order to identify the most appropriate model choice (Breuer et al., 2009; Butts et al., 2004; Donnelly-Makowiecki & Moore, 1999; Duan et al., 2006; Perrin et al., 2001; Reed et al., 2004; Refsgaard & Knudsen, 1996; van der Linden & Woo, 2003; Wood et al., 1998). Depending on the differences in their numerical setup, calibration strategies, or other implementation differences, the obtained results may be limited in their explanatory power (Clark et al., 2011). Modular modeling frameworks have thus been developed to allow controlled experiments and foster the development of multiple working hypotheses for model structure intercomparison and identification. Examples for modular modeling frameworks are MMS.
(Leavesley et al., 2002), RRMT (Wagener et al., 2002), FUSE (Clark et al., 2008), SUPERFLEX (Fenicia et al., 2011), CMF (Kraft et al., 2011), SUMMA (Clark et al., 2015a, 2015b), ECHSE (Kneis, 2015), DECIPHeR (Coxon et al., 2019), and RAVEN (Craig et al., 2020).

Many of these frameworks have been used to conduct model intercomparison studies. The top-down approach starts with a simple model and adds components which are accepted as long as they improve model performance (Fenicia et al., 2016). The bottom-up approach starts with a complex model and removes processes to identify the most significant ones (Chang et al., 2017; Jehn et al., 2018). Others build predefined model structures based on expert knowledge (Gharari et al., 2014; Viglione et al., 2018; Wrede et al., 2015), varying complexity hypotheses (Fenicia et al., 2014; Van Esse et al., 2013) or different combinations of the conceptual choices of existing hydrological models (Coxon et al., 2014; Krueger et al., 2010). Other studies have compared available fixed model structures within common evaluation frameworks (Dams et al., 2015; De Boer-Euser et al., 2017; Knoben et al., 2019; Koch et al., 2016; Nicolle et al., 2014; Orth et al., 2015; Tegegne et al., 2017; Vansteenkiste et al., 2014). The number of simultaneously tested model structures rarely surpasses 20 and only represents a limited, often subjectively chosen part of the possible model space.

A few studies have attempted to expand the available model space for model structure identification. Seiller et al. (2017) introduced an Empirical Multistructure Framework which relies on an “overproduce and select” philosophy aiming to create initial diversity before selecting the best modeling options. They generate a large number of child models by creating various combinations of model structures from 12 lumped conceptual hydrological parent models (up to 108,852 child models including mathematical aberrations) and afterwards select an ensemble of best performing individuals. Their approach aims to reduce the subjectivity in a prior model selection by optimizing commonly shared parameters for all child models and selecting a variable number of best performing ensemble members after calibration.

Chadalawada et al. (2017) used genetic programming to automatically test the available model space of the TANK model (Sugawara, 1979). By treating modeling choices such as the number of tanks utilized, the positions of outlets in the tanks, and the outflow coefficients of the tanks as parameters, they automatically identify the optimal model configuration of the TANK model for two different catchments in the Southern Hemisphere. In this paper, we propose a similar but more general and flexible method to test a predefined model space. It refrains from brute force calibration of a large number of fixed model structures and allows to test various model structure hypotheses differing in the number of processes and/or process algorithms.

The Automatic Model Structure Identification (AMSI) method presented in this paper is the first demonstration of a mixed-integer optimization algorithm to simultaneously optimize model structure choices (integer decision variables) and parameter values (continuous decision variables) in hydrologic modeling. By combining the optimization algorithm with a modular modeling framework, the inclusion of an individual process or process description in the model structure can be treated as an integer (on/off/process choice) variable, while parameters are simultaneously calibrated in a conventional manner. Thus, AMSI is able to sift through a vast number of combinations for a given model and parameter space in order to identify the most suitable model structure for representing the rainfall-runoff behavior of a catchment as evaluated by a specified objective function.

We introduce AMSI in its most reduced form using a lumped conceptual representation of a catchment, even though the concept can, in principle, be applied to a more complex model space with more detailed spatial discretization. By conducting a synthetic and a real-world experiment, we aim to demonstrate the feasibility and suitability of AMSI to infer an adequate model structure for a given catchment based on calibrating against observed discharge.

2. Material and Methods

In this section, we introduce the general concept of AMSI (section 2.1) as well as the tools we used to implement the proposed method, i.e., the hydrologic modeling framework RAVEN (section 2.2) and the mixed-integer implementation of the dynamically dimensioned search (DDS) calibration algorithm (section 2.3). Additionally, the data set (section 2.4) and performance criteria (section 2.5) used for the two benchmark experiments are presented.
2.1. Automatic Model Structure Identification

A simple lumped conceptual model may be abstracted as a set of state variables corresponding to water storages (e.g., soil water and snow), influenced by external forcings (e.g., temperature and precipitation), where water is transferred between water storages and the outlet as dictated by a set of equations describing hydrologic processes (e.g., infiltration and baseflow), while each of these equations requires a set of parameters. Here, we define a time variable state variable vector \( \vec{\psi}(t) \) comprised of \( i = 1, \ldots, N_S \) state variables, a time variable forcing vector \( \vec{F}(t) \) comprised of \( h = 1, \ldots, N_F \) forcings, and a constant parameter vector, \( \vec{P} \), which here includes all parameters which may or may not be used in a given model configuration. A given model is comprised of a set of processes acting upon the state variables. For instance, as in Figure 1, we may consider a hydrologic model consisting of four individual storage compartments (ponded water, soil 1, soil 2, outlet) which are represented as \( \vec{\psi} = \{ \phi_1, \phi_2, \phi_3, \phi_4 \} \).

We also consider four processes \( (g) \) which influence five fluxes \( (f) \) between compartments: runoff/infiltration (which defines the fluxes \( f^1_{1\rightarrow4} \) and \( f^2_{1\rightarrow4} \)), percolation from soil 1 to soil 2 \( (f^3_{2\rightarrow3}) \), capillary rise from soil 2 to soil 1 \( (f^4_{3\rightarrow2}) \), and baseflow \( (f^4_{1\rightarrow4}) \). Note that the superscript \( j \) of each flux \( f \) refers to the process index, while the subscript \( i \rightarrow i^* \) refers to the source \( i \) and destination \( i^* \) storage compartment. The resultant system of equations is given as

\[
\begin{align*}
\frac{\partial \phi_1}{\partial t} & = -f^1_{1\rightarrow4} - f^1_{1\rightarrow2} \\
\frac{\partial \phi_2}{\partial t} & = f^1_{1\rightarrow2} - f^2_{2\rightarrow3} + f^3_{3\rightarrow2} \\
\frac{\partial \phi_3}{\partial t} & = f^2_{2\rightarrow3} - f^3_{3\rightarrow2} - f^4_{3\rightarrow4} \\
\frac{\partial \phi_4}{\partial t} & = f^1_{1\rightarrow4} + f^4_{3\rightarrow4}
\end{align*}
\]

where, quite generally, the fluxes \( f \) are dependent upon the system state, forcings, and parameters, i.e., \( f^j_{i\rightarrowi^*} = f(\vec{\psi}(t), \vec{F}(t), \vec{P}) \). With these four processes, a model structural instance may be considered as defined by the vector of processes \( g^j = \{ g^1, g^2, g^3, g^4 \} \) acting on the state variables, where each component of this process vector refers to the individual model process. For instance, in this example, \( g^1 \), the processes changing the state of ponded water \( \phi_1 \), includes the definitions of both \( f^1_{1\rightarrow4} \) and \( f^1_{1\rightarrow2} \). Each process is associated with an algorithmic description. We can have multiple options for the algorithmic representation of each process and the corresponding fluxes associated with it (i.e., capillary rise \( g^3 \) could be represented using algorithm \( g^3_a \) or \( g^3_b \) or \( g^3_c \) not at all); a specific instance \( k \) of a model is defined by the algorithm choices

\[
\bar{g}^j_k = g^j_k = \{ g^1_{k1} : g^1_{k2} : g^1_{k3} : g^1_{k4} \} = \{ g^2_{k1} : g^2_{k2} : g^2_{k3} : g^2_{k4} \} \]

where \( c_{jk} \) is a choice of a process algorithm for process \( j \) for the model instance \( k \). Optimally selecting model structures therefore corresponds to identifying the most appropriate option \( c_{jk} \) for each process \( g^j \) out of an inevitably finite subset of possible options \( c_j \). The choices \( \bar{c} = \{ c_j \} \) for all processes are the integer decision variables used in model optimization, and we are therefore trying to maximize an objective function by identifying the optimal vectors \( \bar{c}^* \) and \( \bar{P}(\bar{c}^*) \). This may be expressed as the following mixed-integer optimization problem, assuming that the objective function to be optimized is conditional only upon a fit to observed discharges \( Q \):
The optimization algorithm DDS (Tolson & Shoemaker, 2007; Tolson et al., 2009) is combined via a preprocessor with the RAVEN modeling framework, to allow simultaneous model structure and parameter calibration. The preprocessor creates the RAVEN model setup files for a given set of process choices \( \tilde{c} \) and parameter values \( \tilde{P}(\tilde{c}) \) that are selected by DDS as the next candidate solution to test. This setup will hereafter be referred to as the AMSI framework and is described in more detail in section 3.1.

DDS is designed for computationally expensive optimization problems and has been widely used in hydrology (Becker et al., 2019; Seiller et al., 2017; Shafii & Tolson, 2015; Zink et al., 2018). Although Tolson and Shoemaker (2007) originally introduce DDS for solving optimization (model calibration) problems with continuous decision variables, Tolson et al. (2009) modify DDS to solve optimization problems with discrete or integer-valued decision variables. DDS is currently distributed in Matlab (https://uwaterloo.ca/scholar/btelson/software-0) as a mixed-integer optimization algorithm (solves for continuous and integer decision variables simultaneously) where the user specifies the type of each decision variable. This version of DDS is utilized in this research.

\[
[c^*, \tilde{P}(c^*)] = \text{arg} \max_{\tilde{c} \in \tilde{c}} \text{OF}(Q(\tilde{c}, \tilde{P}(\tilde{c}^*)), Q)
\]

where \([c^*, \tilde{P}(c^*)]\) is the optimized structure-parameter combination, \(\tilde{c}\) is the vector set of all model options (the total model space surveyed), \(\tilde{P}_{\text{min}}\) and \(\tilde{P}_{\text{max}}\) are vectors of the minimum and maximum allowable values for each parameter, and \(\text{OF}\) is the objective function, which is a function of the observed discharge \(Q\) and the modeled discharge \(\hat{Q}\). The simultaneous optimization of model structural choices and model parameters is termed AMSI as the identification of a suitable model structure is conducted automatically based on a set of model structural alternatives, the parameter space, and an objective function.

The AMSI framework proposed in this paper combines the modular hydrological modeling framework RAVEN (Craig et al., 2020) and the heuristic, global DDS algorithm (Tolson & Shoemaker, 2007; Tolson et al., 2009) in order to optimize the model construction process. However, AMSI could potentially be any approach that simultaneously optimizes hydrologic model structure and model parameters against one or more performance metrics used as an objective function. We will now introduce the two existing modeling tools the proposed AMSI approach is based on.

### 2.2. The RAVEN Hydrological Modeling Framework

Unlike most hydrological models, the modular modeling framework RAVEN (Craig et al., 2020) is not fixed to a single representation of physical processes describing the function of a watershed. It supports customization of the model structure, watershed delineation, forcing function estimation, and multiple other modeling choices. RAVEN models can thus be simple lumped representations with a couple of conceptual assumptions about the rainfall-runoff behavior of a watershed, or they can be semidistributed with many individual storage compartments using physically based process representations. The complexity of the model is thus limited by the modeler and/or the available data. RAVEN offers an extensive process library to build the model structure the modeler assumes to be adequate for a given watershed but also supports emulation (and augmentation) of a number of existing hydrological models (e.g., the UBC Watershed Model, Quick, 1995; Environment Canada’s version of the HBV model, Bergstrom, 1995; and GR4J, Perrin et al., 2003). In this work, RAVEN is used to build lumped conceptual models running on a daily time step, to reduce the complexity of the presented method.

To define a model structure in RAVEN, the user specifies commands in the model setup files defining the processes to include, the chosen algorithms to use for each process, \(\tilde{c}\), and (if needed) the connections \(\phi\) between the corresponding storage compartments. Thus, changing the structure of the model corresponds to redefining a set of commands in the primary input file. Parameters for RAVEN are also defined in the input files and do not necessarily have to be used by the specified model configuration, which readily supports the conditional use of parameters. RAVEN’s utility for AMSI comes therefore both from its modular design and its support for easily scripting structure and parameter changes.

### 2.3. Mixed-Integer Optimization for AMSI

The optimization algorithm DDS (Tolson & Shoemaker, 2007; Tolson et al., 2009) is combined via a preprocessor with the RAVEN modeling framework, to allow simultaneous model structure and parameter calibration. The preprocessor creates the RAVEN model setup files for a given set of process choices \(\tilde{c}\) and parameter values \(\tilde{P}(\tilde{c})\) that are selected by DDS as the next candidate solution to test. This setup will hereafter be referred to as the AMSI framework and is described in more detail in section 3.1.
Here, the inactivating inactive parameters is a reasonable trade-off because they are not currently active and hence not leading to improved model performance. However, the cost of perturbing only randomly selected decision variables from the current best solution. These perturbation magnitudes are randomly sampled from a normal distribution with a mean of zero. For integer-valued decision variables, the perturbation magnitude is randomly sampled from a discrete probability distribution that approximates a normal distribution (see Tolson et al., 2009 for complete details). The only user-defined algorithmic parameter is the perturbation parameter (\(r\)), which defines the standard deviation of the random perturbation size as a fraction of the decision variable range. The default value of this parameter recommended and used in this study is 0.2. The disregard of parameter sensitivities from sampling the parameter (and model) space is one advantage of using DDS for AMSI as explained in the following example.

In AMSI, the model structural choice vector, \(\tilde{c}\), and all potentially active model parameters, \(\tilde{P}(\tilde{c})\), are combined into one single decision variable vector, and the optimization algorithm (e.g., DDS) must find the optimal values of the decision variables. This allows the simultaneous calibration of model and parameter space and ensures equal treatment of both spaces during the calibration process. The structural choice vector is composed of integer variables, while the model parameters are (usually) continuous variables. The model structure space can be defined as a vector of sets, where each option \(c_j\) (corresponding to process \(j\)) can be assigned a finite number of integer values. For example, the maximum model space for the model defined by Equation 1 can be written as the cartesian product of the process options

\[
\tilde{c} = \{1\} \times \{0, 1\} \times \{0, 1, 2\} \tag{4}
\]

\[
\tilde{c} = \{(1, 1, 0, 0), (1, 1, 0, 1), (1, 1, 0, 2), (1, 1, 1, 0), (1, 1, 1, 1), (1, 1, 1, 2)\} \tag{5}
\]

Here, the infiltration process \((j = 1)\) and percolation process \((j = 2)\) have one algorithmic option, the capillary rise process \((j = 3)\) has one option that may be turned on or off, and the baseflow process may be turned off \((c_4 = 0)\) or use one of two algorithms (linear or nonlinear reservoir). Assuming that capillary rise is not associated with any model parameter while linear baseflow has one parameter \(\theta_1\) and nonlinear baseflow has two parameters \(\theta_1, \theta_2\) and that five additional parameters \(\theta_3\) to \(\theta_7\) are required for the other two (fixed) processes, DDS needs to generate two integer variables \(c_3 \in \{0, 1\}\) and \(c_4 \in \{0, 1, 2\}\), as well as seven continuous variables \(\theta_1\) to \(\theta_7\) in their respective ranges.

Two difficulties arise in this setup: (1) the importance and optimal value of model parameter \(\theta_1\) that exists in both baseflow descriptions might differ depending on which description is active (i.e., if \(c_4 = 1\) or \(c_4 = 2\)). DDS is able to handle this to a certain extent since it ignores the sensitivity of model variables during the sampling of new candidates. (2) DDS is forced to optimize parameters which are not used in the sampled model structure since \(\theta_1\) to \(\theta_7\) are always sampled, independent of the set of processes chosen. This leads to a certain degree of algorithmic inefficiency during the calibration since variables might be perturbed that are not currently active and hence not leading to improved model performance. However, the cost of perturbing inactive parameters is a reasonable trade-off for the ability to handle parameters and process options simultaneously. It should be noted that AMSI does not start the parameter search from scratch every time DDS generates a new candidate structure. AMSI, since it relies on DDS, always perturbs the current best solution to generate a candidate solution. Thus, parameter values associated with the current best model structure are in a sense reused.

### 2.4. MOPEX Data

For testing and applying the AMSI framework, a data set introduced during the second and third Model Parameter Estimation Experiment (MOPEX) workshops (Duan et al., 2006) is used. MOPEX provides data for 12 unregulated basins in the eastern United States which were selected because of their wide range of hydro-climatic conditions. All catchments provide more than 50 years of daily data for mean areal
precipitation, minimum and maximum air temperature, streamflow, potential evaporation as well as soil, land use, and terrain data. Table 1 presents some catchment characteristics highlighting the hydrologic diversity, while Figure 2 shows their location.

The catchments range from very humid (e.g., French Broad and Tygart Valley) to very arid (e.g., Guadalupe and San Marcos) and have been used in many previous studies (Carrillo et al., 2011; Cuntz et al., 2016; Herman et al., 2013; Troch et al., 2013; van Werkhoven et al., 2008). During the MOPEX study (Duan et al., 2006), eight different models of which four (SWB, SAC, GR4J, and PRMS) are rainfall-runoff models and four (ISBA, SWAP, NOAH LSM, and VIC-3L) are land surface models were calibrated for all 12 catchments. These models vary in their complexity from simple conceptual models to complex physically based catchment representations considering Darcy’s and Richards’ equation. Performance results from these eight models as stated in the MOPEX study (Duan et al., 2006) will be compared to the performance of the identified AMSI models in section 4.2.

2.5. Performance Criteria

The choice of objective function for the AMSI framework is very important as the identified model structures will be based solely on the performance regarding the chosen metric. Several studies analyze the value of different hydrological signatures for predicting catchment response (Addor et al., 2018; Euser et al., 2013; McMillan et al., 2017) and are good starting points for identifying a suitable objective function. However, this paper aims to demonstrate AMSI’s feasibility rather than finding the most adequate model for a specific modeling purpose. Thus, the objective function was selected based on more subjective reasoning.

The Nash-Sutcliffe efficiency (NSE) (Nash & Sutcliffe, 1970) is chosen because it is a commonly used metric in hydrology to evaluate the performance of hydrologic models. It allows comparison across catchments as it is a normalized metric that compares the mean squared error (MSE) to the variance (Var) of the observed data. It is also the same metric the MOPEX experiment (Duan et al., 2006) used for calibration and thus allows a better comparison with the MOPEX results in section 4.2. The NSE is defined as follows:

\[
NSE = 1 - \frac{MSE}{Var(Q)} = 1 - \frac{\sum_{i=1}^{N} (q_i - \bar{q}_i)^2}{\sum_{j=1}^{N} (q_j - \bar{Q})^2}
\]

where \(\bar{Q}\) and \(Var(Q)\) are the mean and the variance of the observed flows and \(q_i\) and \(\bar{q}_i\) indicate the observed and simulated data at time step \(i\), respectively.
3. Experimental Design

This section defines the AMSI framework (section 3.1) as well as the experimental design of a synthetic (section 3.2) and a real-world experiment (section 3.3). The synthetic experiment is designed to explore the effectiveness and feasibility of the proposed AMSI approach, while the real-world experiment explores the suitability of the AMSI framework to infer a feasible model structure for a given catchment by calibrating against the observed discharge.

3.1. General AMSI Setup

The proposed AMSI framework is a combination of two existing modeling tools via a preprocessing script. Figure 3 illustrates the general workflow of the AMSI framework. In the first step, the catchment as well as forcing and calibration data needs to be prepared according to RAVEN requirements. The current AMSI framework supports lumped catchment representations, requires at least precipitation and temperature data as climate input, and uses observed discharge as calibration data. The data requirements can vary depending on the processes included in the predefined model space.

The second step requires the definition of the model and parameter space. This includes deciding which processes to include and which process algorithms to test as well as assigning integer parameters to each option. Additionally, the parameter bounds for all possible model parameters need to be specified. The model space $\mathcal{C}$ used for the experiments herein is fixed to include one or two soil storage compartments and 10 different hydrological processes with different process algorithms as defined in Table 2 and depicted in Figure 4a. The number of considered processes and tested process algorithms is readily adjusted. For the infiltration and baseflow process of each soil compartment, four different process algorithms are currently evaluated in the AMSI framework. All other processes can be either on or off and use a fixed process algorithm. This leads to a model space that includes 3,200 different model structures. The AMSI framework allows to fix specific processes or process descriptions in order to include expert or prior knowledge into the tested model space or test isolated model components only. A simple snow routine based on the degree-day method was, for example, fixed for the current AMSI model space and will always be used.

The AMSI framework uses the DDS optimization algorithm to provide a vector with parameter information on how to set up the RAVEN model structure (integer parameters) and how to define the necessary model parameters (continuous parameters). It builds the necessary files and executes the RAVEN model in order to acquire a simulated discharge time series. Using the defined objective function, a performance metric is passed back to DDS to inform the optimization according to Equation 3. After a predefined number of model evaluations, the AMSI framework terminates and provides the optimized structure-parameter combination.
The AMSI framework thus provides an optimized model structure for the representation of the rainfall‐runoff behavior of a catchment according to the chosen objective function. It has to be noted that the number of model evaluations needs to be sufficiently large to extensively search the defined model and parameter space in order to acquire reliable results.

### 3.2. Synthetic Experiment: Reidentification of Model Structures

The synthetic experiment is designed to explore the effectiveness of the proposed AMSI approach. Two sets of experiments are performed. Set A uses a reduced model space which is introduced in addition to the previously introduced full model space to reduce the number of possible model structures for some of the experiments. This allows for an easier analysis of potential process/parameter interactions or compensations. Figure 4b displays the flow diagram of the reduced model space utilized in set A. Set B uses the full model space as introduced in section 3.1 and depicted in Figure 4a. Table 3 provides a short overview of the main differences of both experimental sets.

While the full model space of experimental set B allows for 3,200 unique model structures, the reduced model space of experimental set A allows for 20 unique model structures. It consists of one single soil compartment with two optimizable processes considering eight different process algorithms: four infiltration algorithms and four baseflow algorithms. The baseflow algorithm can be turned off (fifth option), while the infiltration process is always active. Four out of 20 possible model structures would thus have no baseflow if the fifth
option applies. This causes the soil compartment to fill until it spills. These four structures are not considered further for the following reidentification experiments due to the limited realism of such models.

For the remaining 16 model structures of the reduced model space, parameter values are arbitrarily fixed, and the models are used to create a discharge time series based on 5 years of forcing data from the humid French Broad catchment (1995–2000). Each structure is labeled with an experiment identifier (SYN_A 1 to SYN_A 16). Note that French Broad has a high precipitation rate with monthly average values larger than 100 mm and the reduced model space does not consider evaporation. This causes very dynamic synthetic hydrographs if the retention time of the single storage model is low. A 100 mm and the reduced model space does not consider evaporation. This causes very dynamic synthetic hydrographs if the retention time of the single storage model is low. A figure showing the corresponding discharge time series created with the specified models as well as the model structural and parameter choices for each of the 16 experiments in the experimental set A is provided in Figure S1.

All 16 created synthetic discharge time series are then used as calibration data for testing the AMSI framework (based on the reduced model space). The entire 5-year data period was used for calibration including a 90-day warm-up period. The calibration process starts with a randomly generated initial model configuration and parameter set. By calibrating against the synthetic discharge, the original model structure and its corresponding parameters should be identifiable, unless there is equifinality within model structures or parameters. By running 100 trials from different starting points, statistical conclusions about the performance and the potential equifinality within the chosen structure and parameter space can be obtained. The calibration was performed for five different calibration budgets (number of allowed model evaluations) for each of the trials: 500, 1,000, 2,500, 5,000, and 10,000 model evaluations to analyze the efficiency of the complex optimization problem. The NSE is used as objective function.

The same calibration procedure was applied to experimental set B, where four model structures were picked out of the full model space described in section 3.2. The predefined structures and parameters were picked...
arbitrarily and are defined in Table 4. The corresponding synthetic discharge time series are used as calibration goal for AMSI. The AMSI framework was run 100 times with the maximum budget of 10,000 model evaluations to evaluate the reidentification performance for this more complex model space.

3.3. Real-World Experiment: Identifying A Suitable Model Structure via Mixed-Integer Calibration

The real-world experiment is designed to explore the suitability of AMSI to infer a feasible model structure for a given catchment by calibrating against the observed discharge. AMSI will be used on the 12 hydro-climatically differing MOPEX catchments introduced in section 2.4. Based on the full model space defined in section 3.1 AMSI will be applied 100 times (i.e., 100 replicates or optimization trials) for each of the 12 catchments. Such a high number of replicates was chosen to quantify the variability in the identified model structures, which can be an indicator for model structural uncertainty and equifinality. Each replicate has a calibration budget of 10,000 model evaluations.

All catchment models are lumped models and defined by their catchment size, elevation, location, and forest cover. Minimal data requirements are also presumed for the forcing data, consisting of daily data for precipitation, potential evapotranspiration, and maximum and minimum temperature. Measured daily discharge is used for calibration. The calibration period lasts from 1 November 1975 to 31 October 2000 with a 365-day warmup period included. The validation period lasts from 1 November 1950 to 31 October 1975. NSE was chosen as objective function to allow better catchment intercomparison as well as a comparison to the results of the second and third MOPEX workshop (Duan et al., 2006).

In order to pick the most suitable model structure out of the 100 initializations, all model structures identified during calibration are validated. The one performing best in validation is considered the most suitable structure for the catchment. For this structure, an additional fine-tuning parameter calibration with DDS using a budget of 5,000 model evaluations is conducted to get an impression of how much additional model skill can be gained in comparison to the performance of AMSI by additional parameter optimization.

4. Results and Discussion

This section discusses the obtained results for the synthetic (section 4.1) and the real-world experiment (section 4.2).

4.1. Synthetic Experiment: Reidentification of Model Structures

The goal of the synthetic experiments is to reidentify the structures and parameters used to create a specific hydrograph. The experiment was conducted for 16 model structures out of the reduced model space (SYN_A 1 to SYN_A 16) and for four structures out of the full model space (SYN_B 1 to SYN_B 4). The performance regarding the model structure reidentification is shown in Figure 5 for the reduced model space and Figure 6 for the full model space, respectively.

Figure 5 shows in a stacked bar graph how many of the 100 AMSI trials for the reduced experiment identified the true model structure (green), a model structure that differs in the chosen infiltration process (red color spectrum), a model structure that differs in the chosen baseflow process (blue color spectrum), or a model structure that differs in both processes from the true model structure (purple). The results are shown for five different calibration budgets. For the six synthetic experiments SYN_A 3, SYN_A 6, SYN_A 7, SYN_A 8,
SYN_A 11, and SYN_A 12, the AMSI framework is able to correctly identify the model structure in more than 90% of the cases after 5,000 model evaluations. All other experiments show mainly deficits in the correct identification of the infiltration process.

There are several reasons for this. First, the hydrographs are not particularly sensitive to the choice of the infiltration algorithm (see Figure S1). The hydrographs differ clearly when different baseflow processes are used, but only minor differences appear when different infiltration processes are used. This is mainly caused by the simple setup of the synthetic experiments (single conceptual soil storage with large storage size etc.). These minor differences in the hydrograph lead to minimal deviations in an integral diagnostic like the chosen NSE, making the selection of the original process very difficult. This is shown in Figure 7 which plots the NSE for SYN_A 1 to SYN_A 16 of all models with correctly identified baseflow but discriminated by the infiltration algorithm choice. In total, 97.75% of all models have NSE values above or equal to 0.99 and are thus very hard to discriminate. The figure shows the considerable equivalence of the infiltration processes for this setup as the identification of the correct infiltration algorithm was not possible for 10 out of 16 experiments due to identical diagnostic measures. It has to be noted that choosing NSE as an objective function is likely insufficient for structure identification and more process-based indices should be used in future studies.

The reidentification results for the more complex model structures of experimental set B (SYN_B 1 to SYN_B 4) are depicted in Figure 6, which shows a heatmap of all 100 identified AMSI models with each column representing a process and each line representing one model identified using AMSI. The AMSI models (lines) have been clustered by calculating the Euclidean distance to allow a clearer representation of similar identified model structures. Colors are assigned based on the associated integer of each process algorithm as defined in Table 2. Different colors in the same column indicate variability in process identification. Some processes, such as soil evaporation (SEvp) and infiltration (Inf), seem to be well identifiable. The soil evaporation process (SEvp) was identified correctly in 100% of the cases, emphasizing the importance and influence of this process. The infiltration processes was well identifiable except for experiment SYN_B 4 where the GR4J infiltration method was preferred over the correct HBV infiltration method. This could indicate equifinality in the infiltration process. The processes canopy interception and evaporation for snow (CEvpS), capillary rise (CRise), and linear and nonlinear baseflow processes (BF1 + BF2) seem comparatively difficult to identify and show significant variability in most experiments. The exact reasons for the high variability in some processes or experiments are very hard to discriminate but often caused by limited influence on the hydrograph (e.g., CEvpS and CRise), while in other cases process and/or parameters, compensation leads to equifinality in the identified model structures and thus to variability in process choices. Figure 8 shows the effect of equifinality on the reidentification processes. The eight processes are rarely all identified correctly. Most of the time 75% of the identified AMSI models differ in at least two processes from the original model. Nevertheless, all identified AMSI models show NSE values between 0.89 and 1.00 with a median at 0.99. On average, more than 97.5% of the identified AMSI models for experimental set B have an NSE value of 0.95 or higher despite the low reidentification rate. Thus, almost all identified AMSI models are able to reproduce the hydrograph very closely but choose different structure-parameter combinations to do so. In conclusion, the reidentification of model structures is possible but strongly hindered by equifinality in process and parameter choices unique to every model-parameter structure.

Table 3
Details of Two Experimental Sets and the Different Model Spaces Utilized for Each Set

| Model space name | Experimental set A | Experimental set B |
|------------------|--------------------|--------------------|
| Reduced model space | 1 | 2 |
| Full model space | 2 | 10 |
| Soil compartments | 8 | 21 |
| Total number of calibration parameters | 6 | 13 |
| Max. active number of model parameters (per model structure) | 3 | 8 |
| Feasible model structures | 20 | 3,200 |
| Number of experiments | 16 | 4 |

Note. The model spaces differ in the available number of soil compartments, optimizing processes, and parameters. The processes and parameters that are relevant for each model space are indicated in Table 2.

Table 4
Four Fixed Model Structures Used for Experimental Set B Arbitrarily Chosen From the Full Model Space Allowing for 3,200 Model Structures

| Process | SYN_B 1 | SYN_B 2 | SYN_B 3 | SYN_B 4 |
|---------|---------|---------|---------|---------|
| Refrze  | X (-)   | X (-)   | X (-)   |         |
| CEvpR   | X (-)   |         |         |         |
| CEvpS   |         |         |         |         |
| SEvp    | X (-)   | X (-)   |         |         |
| Inf_GR4J|         |         |         |         |
| Inf_PC  |         |         |         |         |
| Inf_HBV |         |         |         |         |
| SR      |         |         |         |         |
| BF1_CON |         |         |         |         |
| BF1_GR4J|         |         |         |         |
| BF1_LIN |         |         |         |         |
| BF1_PL  |         |         |         |         |
| Perc    |         |         |         |         |
| CRise   |         |         |         |         |
| BF2_CON |         |         |         |         |
| BF2_GR4J|         |         |         |         |
| BF2_LIN |         |         |         |         |
| BF2_PL  |         |         |         |         |
| X (0.05) |         |         |         |         |
| X (0.01, 1.5) |         |         |         |         |
| X (10) |         |         |         |         |
| X (10) |         |         |         |         |
| X (35) |         |         |         |         |
| X (1) |         |         |         |         |
| X (5) |         |         |         |         |
| X (250) |         |         |         |         |
| X (0.05) |         |         |         |         |

Note. Parameters were specified for these structures (values in parentheses) and used to derive synthetic hydrographs. These hydrographs are used as ground truth to test if the AMSI framework is able to reidentify these model structures and model parameters when using the full model space of 3,200 model structures and associated parameter ranges. These experiments are labeled SYN_B 1 to SYN_B 4.
Figure 5. Stacked bar graphs for 16 structure reidentification experiments (SYN_A 1 to SYN_A 16) for the reduced model space. Graphs show how many out of 100 AMSI runs identified the true model structure (green), a model structure that differs in the chosen infiltration process (red color spectrum), a model structure that differs in the chosen baseflow process (blue color spectrum), or a model structure that differs in both processes from the true model structure (purple). Five different calibration budgets (500, 1,000, 2,500, 5,000, 10,000) were tested as indicated on the x-axis. The process labels in the legend are explained in Table 2. The underlying set of model structures used for calibration is shown in Figure 4b.

Figure 6. Heatmap of all 100 identified AMSI models for SYN_B 1 to SYN_B 4 with columns representing processes and lines representing single AMSI models. The original model structures that should have been reidentified are defined in Table 4 and depicted in the top row of this figure. The bottom row shows the results for all 100 AMSI models identified with a calibration budget of 10,000 model evaluations. Colors are based on the assigned integer of each process as defined in Table 2. The legend defines how these integer values are translated into on/off, infiltration, and baseflow process options. The underlying set of model structures used for calibration is shown in Figure 4a.
4.2. Real-World Experiment: Identifying A Suitable Model Structure via Mixed-Integer Calibration

By applying AMSI 100 times on each of the 12 MOPEX catchments, several different model structures with varying parameter sets are identified for each catchment. The differences are attributable to different initializations of the optimization run and/or different degrees of convergence toward optimality. Figure 9 shows the calibration and validation results for all 100 identified structure-parameter combinations for each catchment. The number below the catchment label on the x-axis indicates how many unique model structures are part of this 100-member ensemble. Additionally, the calibration results of the second and third MOPEX

Figure 7. NSE values for model structures which identified the correct baseflow, discriminated after the chosen infiltration algorithm (x-axis in each panel). Results are shown for 10,000 model evaluations of the synthetic experiments SYN_A 1 to SYN_A 16 which had the underlying baseflow and infiltration processes indicated by the options highlighted in the gray boxes left and on top of the plot. Note that the y-axis for the bottom line of plots differs.

Figure 8. Histogram of the number of correctly identified processes out of eight possible processes in 100 AMSI trials for the experiments SYN_B 1 to SYN_B 4. This count refers to Figure 6 by counting the number of columns matching the same color as the top row for each of the 100 trials for the four experiments.
workshop as reported by (Duan et al., 2006) for eight different fixed structure models are plotted for comparison. Validation results for the MOPEX models were not available. The catchments are ordered in increasing order based on their mean NSE during calibration, as done in the MOPEX paper. The order of catchments between the AMSI results and the plotted MOPEX results might however slightly differ as no catchment names were given in the MOPEX paper. Validation results for the MOPEX models were not available. Blue dots indicate additional parameter calibration result for the best AMSI validation model using a budget of 5,000 model evaluations. The numbers below the catchment names indicate the number of unique model structures identified during the AMSI runs. Note there is one validation result for catchment GL that is lower than the range displayed. Details about the 12 catchments and their abbreviations can be found in Table 1.

![Figure 9](image)

**Figure 9.** Calibration and validation results for 100 AMSI trials shown as frequency distributions as well as calibration results for eight different models from the second and third MOPEX workshop (Duan et al., 2006) (fig. 11). The underlying model space used for calibration is shown in Figure 4a. The catchments are ordered in increasing order based on their mean NSE during calibration, as done in the MOPEX paper. The order of catchments between the AMSI results and the plotted MOPEX results might however slightly differ as no catchment names were given in the MOPEX paper. Validation results for the MOPEX models were not available. Blue dots indicate additional parameter calibration result for the best AMSI validation model using a budget of 5,000 model evaluations. The numbers below the catchment names indicate the number of unique model structures identified during the AMSI runs. Note there is one validation result for catchment GL that is lower than the range displayed. Details about the 12 catchments and their abbreviations can be found in Table 1.

workshop as reported by (Duan et al., 2006) for eight different fixed structure models are plotted for comparison. Validation results for the MOPEX models were not available. The catchments are ordered in increasing order based on their mean NSE during calibration, as done in the MOPEX paper by (Duan et al., 2006). The order of catchments between the AMSI results and the plotted MOPEX results might however slightly differ as no catchment names were given in the MOPEX paper by (Duan et al., 2006).

![Figure 10](image)

**Figure 10.** (a) Structure of the best performing model in validation for each catchment, selected out of all 100 identified structure-parameter combinations per catchment. The structures are defined by integer values corresponding to specific process descriptions as defined in Table 2. Each column represents a process, each line a catchment. (b) Consistency (%) of the identified best validation model for each catchment from (a), discriminated for each process. Consistency (%) is calculated by dividing the number of models using the same process as the best validation model by the number of all tested models multiplied by 100. Details about the 12 catchments and their abbreviations can be found in Table 1. Details about the processes and their abbreviations can be found in Table 2.
This, however, does not significantly influence the main message of the comparison, which is that the identified AMSI models are almost all within the range of the top 50% of the tested MOPEX models. Considering that some of the MOPEX models are more complex and incorporate additional processes, storages, and data not included in the tested AMSI model space, the AMSI results are considered satisfactory. AMSI is outperforming multiple MOPEX models even with its current toy configuration. It should be noted that the calibration period for the MOPEX models (1980–1998) is a subperiod of the AMSI calibration period (1975–2000).

While most validation results are similar or only slightly below the calibration results (PC, EG, EF, and FB), the results improve for four catchments (BS, RP, MC, and TV). For two catchments (AM and SP), validation results decrease noticeably, while they significantly decrease for the two driest catchments (GL and SM). These catchments have a very low runoff coefficient and in part no flow days. The applied AMSI model space is comparatively limited and thus likely not capable to reproduce this kind of runoff behavior well. An additional parameter calibration for the best models in validation with a budget of 5,000 model evaluations only slightly increases performance, which shows AMSI already identifies acceptable parameter sets for the identified structures. For the arid Guadalupe catchment (GL), however, a significant increase in performance was obtained by additional parameter calibration. This could indicate changing conditions between calibration and validation period, insufficient convergence of the AMSI optimization, and/or an overfitting of an unsuitable model structure. Many studies have shown that arid catchments are more difficult to model than humid catchments (Atkinson et al., 2002; Massmann, 2019) and often require different modeling approaches, e.g., considering processes such as transmission losses (Pilgrim et al., 1988). As the maximum model space currently does not provide this, a low performance is expected for arid catchments.

Regarding the variability in the identified model structures, we can see that there are catchments such as SP, EG, and SM in which many different unique model structures (40, 36, and 34) produce reasonable performance values, while in other catchments such as TV, MC, and PC, a comparatively limited amount of model structures (12, 17, and 18) is identified to work well. No obvious correlation to any catchment characteristics could be identified. To analyze the variability in single process choices, we identify the consistency of a process based on the best performing structure in validation. This is done by selecting the best structure in validation for each catchment as shown in Figure 10a. For each process of this structure, the consistency (%) is calculated by dividing the number of models using the same process by the number of all tested models, multiplied by 100. Each process is analyzed individually for each catchment. If the same process is selected repeatedly, a high consistency is associated to this process. If the consistency is low, the best validation structure differs significantly from other identified model structures in this process. Reasons for a low consistency can be equifinality in process choice, equifinality due to parameter and/or process compensation or the insignificance of the process for the representation of the hydrograph.

Figure 10 shows that every identified AMSI model for each of the 12 catchments identified the soil evaporation process (SEvp) as important. Expectedly, models with the soil evaporation process turned on were consistently better than models without a soil evaporation process. The consistency of this process is thus 100% in all catchments. Another process that can be identified with a relative high consistency is the refreeze process (Refrze). It is turned off in eight of 12 catchments with a consistency of 100%. Surprisingly, the arid catchments San Marcos (SM) and Guadalupe (GL) identify the refreeze process as necessary with a consistency of 82% or as unnecessary with a consistency of 43%, despite a snow rate of less than 1% (Cuntz et al., 2016). Similarly, both catchments show a medium consistency of 42% and 57% for the Canopy Interception and Evapotranspiration of Snow (CEvpS), indicating that this process is turned off in just about half of all identified AMSI models. This suggests insensitivity of the objective function to this process and indicates that snow processes do not have a relevant impact on the representation of the hydrograph in these catchments. It does not seem to matter if they are turned on or off. It can be concluded that a consistency between 40% and 60% suggests insensitivity of a two-choice process. In the case of the San Marcos (SM) catchment, the reason for selecting the refreeze process with a consistency of more than 80% is however unclear. It is likely that due to the simple catchment and calibration setup (lumped catchments calibrated for discharge only), the identified model structures are in part right for the wrong reasons as stated by Kirchner (2006). The current AMSI-framework must thus be considered more of a data fitting experiment. However, the potential of the AMSI method for improving the physical realism of the identified model structures is given by the possibility to include more data and a more rigorous calibration routine (as in using several relevant signatures).
Analyzing the consistency of the canopy interception and evaporation process for rain (CEvpR), we can assume that this is a relevant process for all catchments as it is always turned on, except for Amite River (AM) where the consistency of this choice is just 5%. The consistency of this process in the French Broad (FB) catchment is close to 50% indicating irrelevance of the process. This is feasible as French Broad has one of the highest precipitation rates, causing interception to become less important in the overall water balance. Generally, lower consistency values are associated with the processes infiltration (Inf) and baseflow (BF1 and BF2). This is mainly caused not only by equifinality but also in part by the fact that these processes have several process choices. If just two choices perform equally well (e.g., a linear baseflow and a power law baseflow with an exponent parameter close to one), the consistency will drop significantly. The analysis of reasons causing the variability/equifinality in process choice requires a detailed analysis of each individual model and is considered to be out of the scope of this paper.

Despite the variability in the identified model structures, the AMSI framework was capable of identifying feasible model structures and avoiding infeasible model structures for all 12 MOPEX catchments. The automatically identified model structures perform reasonably well during the validation period and are able to sufficiently reproduce the rainfall-runoff behavior of the catchments as compared through direct evaluation against the eight MOPEX models. For an excerpt of the calibrated and observed discharge time series of each catchment, Figure S2 is provided.

5. Summary and Conclusions

The impact of model structure on model results is a generally acknowledged fact in the hydrological modeling literature. Nevertheless, a lot of time and computer power is invested into identifying suitable model parameters, whereas the selection of (a) suitable model structure(s) often remains an optional step in the modeling chain. This study introduced an approach for AMSI which allows for the simultaneous calibration of model structure and model parameters by combining the modular modeling framework RAVEN with the mixed-integer DDS optimization algorithm. The proposed AMSI framework proved to be capable of identifying feasible model structures and avoiding infeasible model structures for 12 hydro-climatically differing MOPEX catchments by simply calibrating against discharge. We were able to show that all identified AMSI models perform within the range of the top 50% of the eight tested MOPEX models (Duan et al., 2006) during calibration; these tested models are for the most part much more complex than the models created from the AMSI model space. Further experiments investigated equifinality in a predefined model space by attempting to reidentify given model structures from a hydrograph alone. We found that reidentification is possible but revealed considerable variability in the selected model structures due to parameter and process compensation as well as the insensitivity of the hydrograph to certain processes or process descriptions. The novelty of the AMSI approach lies in the automatized testing of different model structure hypothesis and the easily adjustable size and complexity of the testable model and parameter space. We hope that the availability of an automatized approach will foster the testing of different structural hypothesis in future modeling studies. This would in turn reduce the subjectivity generally connected with model selection.

The experiments conducted here showed a few general issues with the current implementation of AMSI. First, the insensitivity of the hydrograph to different process descriptions or processes leads to a high variance in identified model structures. If the impact of a specific process is not measurable by the chosen calibration metric and/or data, AMSI will not be able to decide if a process is relevant or if one process description is more suitable than another. Second, AMSI can lead to model structures that are right for the wrong reasons if it is not supported by additional calibration data or suitable calibration metrics. Third, parameter and process compensation lead to a high variance in identified model structures due to near equivalent diagnostic measures for multiple structure-parameter combinations (equifinality). Lastly, the potential for inefficiency of the current optimization setup mostly due to the presence of conditionally inactive parameters is acknowledged. This complication may lead to DDS finding nonoptimal or only locally optimal model configurations, though there was limited evidence of this in the particular experiments presented here. The cost of perturbing inactive parameters is currently considered a reasonable trade-off for the ability to handle a nonstatic number of parameters, but a more effective approach may include modifying the DDS algorithm to more intelligently support conditionally inactive parameters.
Most of the above issues call for a careful selection and an extension of the utilized calibration metrics and suggest lines of future research that may further improve our ability to automatically identify optimal structure and parameters of hydrologic models. It is expected that the inclusion of alternate data sources or the use of hydrologic signatures on the AMSI objective function is likely to aid in reducing process model equifinality. Future work with AMSI should thus consider revised calibration formulations utilizing hydrologic signatures, case studies with multiple types of system response data, and thus mixed-integer multi-objective optimization algorithms. This would allow for large sample studies on process choices and process sensitivities for hydro-climatically distinct catchments. We also believe it would be worthwhile to identify more efficient ways of calibrating a nonstatic number of parameters to reduce the cost of calibration. Standard calibration algorithms could be adjusted to allow conditional parameter calibration if provided with information how parameters and process descriptions are connected. Meanwhile, utilizing expert knowledge to limit the model and parameter space may improve the identification performance especially for single catchment studies. Equifinality, however, will continue to be a part of hydrological modeling as long as missing data lead to so called ill-posed calibration problems (Beven, 2006). The benefit of AMSI in this regard is an extended exploration of the available model and parameter space than previously possible, even if the model space remains necessarily limited and subjectivity will still be introduced by decisions on model configurations as described by Melsen et al. (2019).

Data Availability Statement

The MOPEX data are available at https://hydrology.nws.noaa.gov/pub/gcip/mopex/US_Data/. The RAVEN model 2.8.1 used in this study and its newer version are available at https://raven.uwaterloo.ca/. DDS 1.2 is available at https://uwaterloo.ca/scholar/btolson/software-0. All data for the analyses in this study are publicly available online at https://www.hydroshare.org/resource/7ee0d33e0cb4e91abe3173d71b3395.

References

Addor, N., & Melsen, L. A. (2019). Legacy, Rather Than Adequacy, Drives the Selection of Hydrological Models. Water Resources Research, 55, 378–390. https://doi.org/10.1029/2018WR022958

Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N., & Clark, M. P. (2018). A ranking of hydrological signatures based on their predictability in space. Water Resources Research, 54, 8792–8812. https://doi.org/10.1029/2018WR022606

Atkinson, S. E., Woods, R. A., & Shpitalnik, M. (2002). Climate and landscape controls on water balance model complexity over changing timescales. Water Resources Research, 38(12), 501–501–17. https://doi.org/10.1029/2002wr001487

Becker, R., Koppa, A., Schulz, S., Usman, M., aus der Beek, T., & Schth, C. (2019). Spatially distributed model calibration of a highly managed hydrological system using remote sensing-derived ET data. Journal of Hydrology, 577, 123,944. https://doi.org/10.1016/j.jhydrol.2019.123944

Bergstrom, S. (1995). Computer models of watershed hydrology. pp. 443–476; Water Resources Publications.

Beven, K. (2001). On hypothesis testing in hydrology. Hydrological Processes, 15(9), 1655–1657. https://doi.org/10.1002/hyp.436

Beven, K. (2006). A manifesto for the equifinality thesis. Journal of Hydrology, 320(1–2), 18–36. https://doi.org/10.1016/j.jhydrol.2005.07.007

Beven, K. J. (2018). On hypothesis testing in hydrology: Why falsification of models is still a really good idea. Wiley Interdisciplinary Reviews: Water, 5(3), e1278. https://doi.org/10.1002/wat2.1278

Breuer, L., Huisman, J. A., Willems, P., Bormann, H., Bronstert, A., Croke, B. F. W., et al. (2009). Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM). I: Model intercomparison with current land use. Advances in Water Resources, 32(2), 129–146. https://doi.org/10.1016/j.adwatres.2008.10.003

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

Carrillo, G., Turchin, P. A., Sivapalan, M., Wagener, T., Harman, C., & Savic, W. (2011). Catchment classification: Hydrological analysis of catchment behavior through process-based modeling along a climate gradient. Hydrology and Earth System Sciences, 15(11), 3411–3430. https://doi.org/10.5194/hess-15-3411-2011

Chadalawada, J., Hlavicek, V., & Babovic, V. (2017). A genetic programming approach to system identification of rainfall-runoff models. Water Resources Management, 31, 3975–3992. https://doi.org/10.1007/s11269-017-1719-1

Chamberlin, T. C. (1965). The method of multiple working hypotheses. Science, 148(3671), 754–759.

Chang, Y., Wu, J., Jiang, G., & Kang, Z. (2017). Identification of the dominant hydrological process and appropriate model structure of a karst catchment through stepwise simplification of a complex conceptual model. Journal of Hydrology, 548, 75–87. https://doi.org/10.1016/j.jhydrol.2017.02.050

Clark, M. P., Kavetski, D., & Fenicia, F. (2011). Pursuing the method of multiple working hypotheses for hydrological modeling. Water Resources Research, 47, W09301. https://doi.org/10.1029/2010WR009827

Clark, M. P., Nijssen, B., Landerquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., et al. (2015a). A unified approach for process based hydrologic modeling 1: Modeling concept. Water Resources Research, 51, 2498–2514. https://doi.org/10.1002/2015WR017198

Clark, M. P., Nijssen, B., Landerquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., et al. (2015b). A unified approach for process based hydrologic modeling: 2. Model implementation and case studies. Water Resources Research, 51, 2515–2542. https://doi.org/10.1002/2015WR017200
Clark, M. P., Slater, A. G., Rupg, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., et al. (2008). Framework for understanding structural errors (FUSE): A modular framework to diagnose differences between hydrological models. Water Resources Research, 44, W00B02. https://doi.org/10.1029/2007WR006735

Coxon, G., Freer, J., Lane, R., Dunne, T., Knoben, W. J. M., Howden, N. J. K., et al. (2019). DECIPHER v1: Dynamic fluxes and Connectivity for Predictions of Hydrology. Geoscientific Model Development, 12(6), 2285–2306. https://doi.org/10.5194/gmd-12-2285-2019

Coxon, G., Freer, J., Wagener, T., Odoni, N. A., & Clark, M. (2014). Diagnostic evaluation of multiple hypotheses of hydrological behaviour in a limits-of-acceptability framework for 24 UK catchments. Hydrological Processes, 28(25), 6135–6150. https://doi.org/10.1002/hyp.10096

Craig, J. R. (2018). Raven user’s and developer’s manual (version 2.8.1).

Craig, J. R., Brown, G., Chlumsky, R., Jenkinson, R. W., Jost, G., Lee, K., et al. (2020). Flexible watershed simulation with the Raven hydrological modelling framework. Environmental Modelling & Software, 129(December 2019), 104,728. https://doi.org/10.1016/j.envsoft.2020.104728

Cutmz, M., Mai, S., Samaniego, L., Clark, M., Wulfmeyer, V., Branch, O., et al. (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, 10,676–10,700. https://doi.org/10.1002/2016JD025097

Dams, J., Nossent, J., Senbeta, T. B., Willems, P., & Batelaan, O. (2015). Multi-model approach to assess the impact of climate change on runoff. Journal of Hydrology, 529, 1601–1616. https://doi.org/10.1016/j.jhydrol.2015.08.023

De Boer-Euser, T., Bouaziz, L., De Nieël, J., Brauer, C., Dewals, B., Drooge, G., et al. (2017). Looking beyond general metrics for model comparison—lessons from an international model intercomparison study. Hydrology and Earth System Sciences, 21(1), 423–440. https://doi.org/10.5194/hess-21-423-2017

Di Baldassarre, G., & Montanari, A. (2009). Uncertainty in river discharge observations: A quantitative analysis. Hydrology and Earth System Sciences, 13(6), 931–939. https://doi.org/10.5194/hess-13-913-2009

Donnelly-Makowiecki, L., & Moore, R. (1999). Hierarchical testing of three rainfall runoff models in small forested catchments. Journal of Hydrology, 239(3–4), 136–152. https://doi.org/10.1016/S0022-1694(99)00056-6

Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H. V., et al. (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. https://doi.org/10.1016/j.jhydrol.2005.07.031

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., & Savenije, H. H. (2013). A framework to assess the realism of model structures using hydrological signatures. Hydrology and Earth System Sciences, 17(5), 1893–1912. https://doi.org/10.5194/hess-17-1893-2013

Fenicia, F., Kavetski, D., & Savenije, H. H. G. (2011). Elements of a conceptual model representation: Is there a correspondence? Hydrologia, 68, 110–121. https://doi.org/10.1007/s13345-010-0025-8

Fenicia, F., Kavetski, D., Hrachowitz, M. H. G., Clark, M. P., Schoups, G., & Freer, J. (2014). Catchment properties, function, and conceptual model representation: Is there a correspondence? Hydrological Processes, 28(4), 2451–2467. https://doi.org/10.1002/hyp.9732

Fenicia, F., Kavetski, D., Savenije, H. J., & Pfister, L. (2016). From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions. Water Resources Research, 52, 954–989. https://doi.org/10.1002/2015WR017398

Gharahi, S., Hrachowitz, M., Fenicia, F., Gao, H., & Savenije, H. H. (2014). Using expert knowledge to increase realism in environmental system models can dramatically reduce the need for calibration. Hydrology and Earth System Sciences, 18(12), 4839–4859. https://doi.org/10.5194/hess-18-4839-2014

Hereman, J. D., Reed, P. M., & Wagener, T. (2013). Time-varying sensitivity analysis clarifies the effects of watershed model formulation on model behavior. Water Resources Research, 49, 1400–1414. https://doi.org/10.1002/wrcr.20124

Jehn, F. U., Breuer, L., Houaska, T., Bestian, K., & Kraft, P. (2018). Incremental model breakdown to assess the multi-hypotheses problem. Hydrology and Earth System Sciences, 22(8), 4565–4581. https://doi.org/10.5194/hess-22-4565-2018

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. Water Resources Research, 42, W03S04. https://doi.org/10.1029/2005WR004362

Kneale, D. (2015). A lightweight framework for rapid development of object-based hydrological model engines. Environmental Modelling & Software, 68, 110–121. https://doi.org/10.1016/j.envsoft.2015.02.009

Knoben, W. J., Freer, J. E., Fowler, K. J., Peel, M. C., & Woods, R. A. (2019). Modular assessment of rainfall-runoff models toolbox (MARRMoT) v1.2: An open-source, extendable framework providing implementations of 46 conceptual hydrological models as continuous state-space formulations. Geoscientific Model Development, 12(6), 2463–2480. https://doi.org/10.5194/gmd-12-2463-2019

Koch, J., Cornelissen, T., Fang, Z., Bogaen, H., Dieckkrüger, B., Kollet, S., & Stisen, S. (2016). Inter-comparison of three distributed hydrological models with respect to seasonal variability of soil moisture patterns at a small forested catchment. Journal of Hydrology, 533, 234–249. https://doi.org/10.1016/j.jhydrol.2015.12.002

Kraft, P., Vaché, K. B., Frede, H.-G., & Breuer, L. (2001). A hydrological programming language extension for integrated catchment models. Environmental Modelling & Software, 26(6), 828–830. https://doi.org/10.1016/j.envsoft.2010.12.009

Kruiger, T., Freer, J., Quinton, J. N., Maclowd, C. J. A., Bilotta, G. S., Brazier, R. E., et al. (2010). Ensemble evaluation of hydrological model hypotheses. Water Resources Research, 46, W07516. https://doi.org/10.1029/2009WR007845

Leavesley, G. H., Markstrom, S. L., Restrepo, P. J., & Viger, R. J. (2002). A modular approach to addressing model design, scale, and parameter estimation issues in distributed hydrological modelling. Hydrological Processes, 16(2), 173–187. https://doi.org/10.1002/hyp.344

Massmann, C. (2019). Identification of factors influencing hyporadric model performance using a top down approach in a large number of U.S. catchments. Hydrological Processes, 34, 4–20. https://doi.org/10.1002/hyp.13566

McMillan, H., Westerberg, I., & Branger, F. (2017). Five guidelines for selecting hydrological signatures. Hydrological Processes, 31(26), 4757–4761. https://doi.org/10.1002/hyp.11300

Melsen, L. A., Teuling, A. J., Törès, P. J., Zappa, M., Mizukami, N., Mendoza, P. A., et al. (2019). Subjective modeling decisions can significantly impact the simulation of flood and drought events. Journal of Hydrology, 568(September 2017), 1093–1104. https://doi.org/10.1016/j.jhydrol.2018.11.046

Montanari, A., Shoemaker, C. A., & van de Giesen, N. (2009). Introduction to special section on uncertainty assessment in surface and subsurface hydrology: An overview of issues and challenges. Water Resources Research, 45, W00B08. https://doi.org/10.1029/2009WR008471

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models: Part I—A discussion of principles. Journal of Hydrology, 10, 282–290.
Nicoll, P., Pushpalatha, R., Perrin, C., François, D., Thiéry, D., Mathevet, T., et al. (2014). Benchmarking hydrological models for low-flow simulation and forecasting on French catchments. *Hydrology and Earth System Sciences*, 18(8), 2829–2857. https://doi.org/10.5194/hess-18-2829-2014

Orth, R., Staudinger, M., Seneviratne, S. I., Seibert, J., & Zappa, M. (2015). Does model performance improve with complexity? A case study with three hydrological models. *Journal of Hydrology*, 523, 147–159. https://doi.org/10.1016/j.jhydrol.2015.01.044

Perrin, C., Michel, C., & Andréassian, V. (2001). Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *Journal of Hydrology*, 242(3–4), 275–301. https://doi.org/10.1016/S0022-1694(00)00393-0

Perrin, C., Michel, C., & Andréassian, V. (2003). Improvement of a parsimonious model for streamflow simulation. *Journal of Hydrology*, 279(1–4), 275–289. https://doi.org/10.1016/S0022-1694(03)00225-7

Pilgrim, D. H., Chapman, T. G., & Doran, D. G. (1988). Problèmes de la mise au point de modèles pluie

Refsgaard, J. C., & Knudsen, J. (1996). Operational validation and intercomparison of different types of hydrological models. *Water Resources Research*, 32(7), 2189–2202. https://doi.org/10.1029/96WR00896

Seiler, G., Ancill, F., & Roy, R. (2017). Design and experimentation of an empirical multifracture framework for accurate, sharp and reliable hydrological ensembles. *Hydrology and Earth System Sciences*, 525, 313–340. https://doi.org/10.1016/j.hydrol.2017.07.002

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

Sugawara, M. (1979). Automatic calibration of the tank model / l’étalonnage automatique d’un modèle à cisterne. *Hydrological Sciences Journal*, 33(4), 379–400. https://doi.org/10.1080/02666668509491261

Quick, M. (1995). Computer models of watershed hydrology, pp. 233–260: Water Resources Publications.

Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D.-J., & DMIF Participants(2004). Overall distributed model intercomparison project results. *Journal of Hydrology*, 298(1–4), 27–60. https://doi.org/10.1016/j.jhydrol.2004.03.031

Refsgaard, J. C., & Knudsen, J. (1996). Operational validation and intercomparison of different types of hydrological models. *Water Resources Research*, 32(7), 2189–2202. https://doi.org/10.1029/96WR00896

Tegegne, G., Park, D. K., & Kim, Y. O. (2017). Comparison of hydrological models for the assessment of water resources in a data-scarce region, the Upper Blue Nile River Basin. *Journal of Hydrology: Regional Studies, 14* (November), 49–66. https://doi.org/10.1016/j.jhrsr.2017.10.002

Tolson, B. A., Asadzadeh, M., Maier, H. R., & Zecchin, A. (2009). Hybrid discrete dynamically dimensioned search (HD-DDS) algorithm for water distribution system design optimization. *Water Resources Research*, 45, W12416. https://doi.org/10.1029/2008WR007673

Tolson, B. A., & Shoemaker, C. A. (2007). Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resources Research*, 43, W01413. https://doi.org/10.1029/2005wr004723

Trost, P. A., Carrillo, G., Sivapalan, M., Wagener, T., & Sawicz, K. (2013). Climate-vegetation-soil interactions and long-term hydrologic partitioning: Signatures of catchment co-evolution. *Hydrology and Earth System Sciences, 17*(6), 2209–2217. https://doi.org/10.5194/hess-17-2209-2013

Van Esse, W. R., Perrin, C., Booij, M. J., Augustijn, D. C., Fénica, F., Kavetski, D., & Lobligeois, F. (2013). The influence of conceptual model structure on model performance: A comparative study for 237 French catchments. *Hydrology and Earth System Sciences, 17*(10), 4227–4239. https://doi.org/10.5194/hess-17-4227-2013

van Werkhoven, K., Wagener, T., Reed, P., & Tang, Y. (2008). Characterization of watershed model behavior across a hydroclimatic gradient. *Water Resources Research, 44*, W01429. https://doi.org/10.1029/2007WR006271

van der Linden, S., & Woo, M.-K. (2003). Application of hydrological models with increasing complexity to subarctic catchments. *Journal of Hydrology*, 270(1–2), 145–157. https://doi.org/10.1016/S0022-1694(02)00291-3

Vansteenkiste, T., Tavakoli, M., Van Steenbergen, N., De Smidt, F., Batelaan, O., Pereira, F., & Willems, P. (2014). Intercomparison of five lumped and distributed models for catchment runoff and extreme flow simulation. *Journal of Hydrology*, 511, 335–349. https://doi.org/10.1016/j.jhydrol.2014.01.050

Viglione, A., Rogger, M., Pirkli, H., Parajka, J., & Blöschl, G. (2018). Conceptual model building inspired by field-mapped runoff generation mechanisms. *Journal of Hydrology and Hydromechanics, 66*(3), 303–315. https://doi.org/10.2478/jhhy-2018-0010

Wagener, T., Lees, M. J., & Wheeler, H. S. (2002). A toolkit for the development and application of parsimonious hydrological models. In Singh, F.-M. (Ed.), *Mathematical models of small watershed hydrology - Volume 2*: Water Resources Publications LLC.

Wood, E. F., Lettenmaier, D. P., Liang, X., Loehman, D., Boone, A., Chang, S., et al. (1998). The project for intercomparison of land-surface parameterization schemes (PILPS) phase 2(c) Red-Arkansas River basin experiment: 1. Experiment description and summary inter-comparisons. *Global and Planetary Change, 19* (1–4), 115–135. https://doi.org/10.1016/S0921-8181(98)00044-7

Wrede, S., Fénica, F., Martinez-Carreras, N., Juilleret, J., Hissler, C., Krein, A., et al. (2015). Towards more systematic perceptual model development: A case study using 3 Luxembourgish catchments. *Hydrological Processes, 29*(12), 2731–2750. https://doi.org/10.1002/hyp.10393

Zink, M., Mai, J., Cuntz, M., & Samaniego, I. (2018). Conditioning a hydrologic model using patterns of remotely sensed land surface temperature. *Water Resources Research, 54*, 2976–2998. https://doi.org/10.1002/2017WR021346