Water Resources Research

RESEARCH ARTICLE
10.1029/2019WR025975

A Brief Analysis of Conceptual Model Structure Uncertainty Using 36 Models and 559 Catchments

W. J. M. Knoben[1], J. E. Freer[2,3], M. C. Peel[4], K. J. A. Fowler[4], and R. A. Woods[1]

[1]Department of Civil Engineering, University of Bristol, Bristol, UK, [2]Now at University of Saskatchewan Coldwater Laboratory, Canmore, Alberta, Canada, [3]School of Geographical Science, University of Bristol, Bristol, UK, [4]Department of Infrastructure Engineering, University of Melbourne, Melbourne, Victoria, Australia

Key Points:
- Conceptual model structure uncertainty is high across different catchments and objective functions
- There is no evidence of systematic overfitting for models with up to 15 calibrated parameters
- Model performance relates more to streamflow signatures than to climate or catchment descriptors

Supporting Information:
- Supporting Information S1

Correspondence to:
W. J. M. Knoben,
wouter.knoben@usask.ca

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

Received 17 JUL 2019
Accepted 29 JUN 2020
Accepted article online 6 JUL 2020

Abstract The choice of hydrological model structure, that is, a model's selection of states and fluxes and the equations used to describe them, strongly controls model performance and realism. This work investigates differences in performance of 36 lumped conceptual model structures calibrated to and evaluated on daily streamflow data in 559 catchments across the United States. Model performance is compared against a benchmark that accounts for the seasonality of flows in each catchment. We find that our model ensemble struggles to beat the benchmark in snow-dominated catchments. In most other catchments model structure equifinality (i.e., cases where different models achieve similar high efficiency scores) can be very high. We find no relation between the number of model parameters and performance during either calibration or evaluation periods nor evidence of increased risk of overfitting for models with more parameters. Instead, the choice of model parametrization (i.e., which equations are used and how parameters are used within them) dictates the model's strengths and weaknesses. Results suggest that certain model structures are inherently better suited for certain objective functions and thus for certain study purposes. We find no clear relationships between the catchments where any model performs well and descriptors of those catchments' geology, topography, soil, and vegetation characteristics. Instead, model suitability seems to relate strongest to the streamflow regime each catchment generates, and we have formulated several tentative hypotheses that relate commonalities in model structure to similarities in model performance. Modeling results are made publicly available for further investigation.

1. Introduction

There is an ongoing debate in hydrology whether a "one model fits all" approach should be pursued, based on the assumption that the fundamental hydrological processes are the same everywhere (e.g., Fenicia et al., 2011; Linsley, 1982; Perrin et al., 2003; Savenije, 2009). This assumption has led to development of rainfall runoff models that are designed to be applied across a wide range of catchments (see, e.g., discussion of the GR4J model in Fenicia et al., 2011, and consider more recent applications of this model in 142 catchments in the United States Oudin et al., 2018). This assumption is contrasted by the concept of "uniqueness of place" (Beven, 2000), the idea that in a practical sense every catchment is unique because there are limits to our understanding of fundamental processes and the availability of sufficiently detailed measurements. As a result of this uniqueness, many hydrological models have been developed that all aim to represent the dominant processes in a given catchment (e.g., Singh & Woolhiser, 2002). While theoretically we should be able to use a single model based on fundamental hydrologic principles, in practice there are many different models available that all represent a certain view of which hydrologic processes are important and how these should be mathematically represented. Choosing an appropriate model out of all possible options is critical to obtain accurate simulations that are the result of plausible representations of the hydrology in a given catchment (Kirchner, 2006). Knowing how much uncertainty is associated with the choice of model structure is also important for quantifying the reliability of model predictions (e.g., Biondi et al., 2012).

Conceptual hydrologic models are the focus of this study. Many different conceptual models exist, and there is much variety in how these models work. Models such as GR4J (four parameters Perrin et al., 2003) use a process-aggregated approach where model fluxes represent the aggregated results of all possible processes. Others such as MODHYDROLOG (15 parameters Chiew, 1990; Chiew & McMahon, 1994) follow a more process-explicit approach where fluxes and states explicitly relate to specific processes such as interception,

infiltration, surface storage, and groundwater-channel exchanges. Many models are somewhere in between and combine explicit process representation with more aggregated approaches. The choice of model structure is one of the main sources of uncertainty in a modeling study (e.g., Andréassian et al., 2009; Coron et al., 2012; Fenicia et al., 2008, 2014; Krueger et al., 2010a; Van Esse et al., 2013), but between-model differences are currently not well understood (Ceola et al., 2015; Gupta et al., 2012). This is of particular concern when models are applied in large numbers of catchments, in which case multiple models might be plausible representations of hydrologic behavior, but it is unsure which of the available models is the most appropriate choice in any given catchment.

Large-sample modeling studies can increase our understanding of model functioning, and the need for them has been discussed often (e.g., Addor et al., 2019; Andréassian et al., 2009; Coxon et al., 2019; Franchini & Pacciani, 1991; Hrachowitz & Clark, 2017; Lane et al., 2019; Linsley, 1982; Perrin et al., 2001; Seiller et al., 2012; Sittner, 1976). Assessing a single model's performance under a wide range of different conditions can lead to increased understanding of the model's strengths and weaknesses. Comparing the performance of different models for a given catchment can lead to increased understanding of hydrologic similarity and between-model differences. Past studies have often needed to limit either the number of catchments or the number of models (see, e.g., Bell et al., 2001; de Boer-Euser et al., 2017; Euser et al., 2013; Franchini & Pacciani, 1991; Krueger et al., 2010b; Lane et al., 2019; Lidén & Harlin, 2000; Moore & Bell, 2001; Nijzink et al., 2016; Perrin et al., 2001; Seiller et al., 2012; Van Esse et al., 2013). This has often been due to limitations in computing power or available data or both. Additionally, the model code that generated such results is often not publicly available, limiting the reproducibility and transparency of such work. With increasing computing power, the availability of open-source model intercomparison frameworks (e.g., Clark et al., 2015; Knoben, Freer, Fowler, et al., 2019; Kraft et al., 2011), and new publicly available data sets for large-sample hydrology (e.g., Addor et al., 2017; Alvarez-Garreton et al., 2018), the concerns that have limited large-sample modeling studies in the past have become somewhat less critical.

Given the current incomplete knowledge on between-model differences, an opportunity now exists to study similarities and differences in the behavior of multiple different models across a wide variety of places. A challenge of such a study is that it can be difficult to keep analysis and visualization manageable due to the large number of results involved. Investigating every interesting individual case is infeasible, and instead lessons must be learned from emergent patterns across the full sample (Hrachowitz & Clark, 2017). Large-sample emergent patterns can provide unique insights into how well models function across a variety of different catchment types and inform understanding of hydrologic similarity between different places. Potential benefits of such studies include more thorough understanding of the places where a given hydrologic model can be used with some measure of confidence and improved ability to model ungauged catchments through regionalization approaches. The aim of this paper is thus to explore the performance and associated model structure uncertainty of 36 conceptual hydrologic models across 559 catchments, covering a wide range of climatic and catchment conditions. Our research objectives are further specified in section 2.

2. Rationale, Research Questions, and Methodology

In this study, we calibrate 36 different model structures for streamflow simulations in 559 catchments using three different objective functions and evaluate model performance during a separate time period (details in section 3). This gives a total sample of 60,372 model application test cases. This section defines four research questions and describes how the modeling results are analyzed to answer these questions.

2.1. Defining a Lower Level of Expected Model Performance

Our approach to model calibration and evaluation expresses model performance as Kling-Gupta efficiency scores (KGE Gupta et al., 2009). A score of 1 indicates perfect agreement between simulations and observations. Scores lower than 1 are difficult to interpret beyond "higher is better," and KGE does not include a built-in benchmark that can be used to distinguish "good" and "bad" scores (Gupta et al., 2009; Knoben et al., 2019). Therefore, we first specify a lower benchmark that provides the necessary context to interpret model KGE scores (Garrick et al., 1978; Pappenberger et al., 2015; Schaefli & Gupta, 2007; Seibert, 2001; Seibert et al., 2018). The lower benchmark is the minimum score we expect each model to obtain before we

consider the model a plausible choice for the catchment under consideration. Research Question 1 is thus *What level of model performance do we expect any model to obtain before we consider the model a plausible option for a given catchment?*

A traditional benchmark (i.e., the "score to beat") in hydrology is the mean annual flow due to its inclusion in the Nash-Sutcliffe efficiency score (NSE Nash & Sutcliffe, 1970). This choice is both quite simplistic (e.g., Garrick et al., 1978) and not equally difficult to beat in different catchments (Schaefli & Gupta, 2007), depending on how seasonally variable the flow in any given catchment is. The interannual mean for every calendar day has been proposed as a benchmark that can account for seasonality in the flow regime (Garrick et al., 1978; Schaefli & Gupta, 2007), provided that this seasonality is stable between years. This is not the case in catchments where the flow observations on any given calendar day are heavily skewed, as might be the case in catchments with very irregular occurrences of high flow peaks. In such cases the interannual mean might be far away from many of the sample values and the interannual median will be closer and more representative of the typical flow regime.

We therefore calculate both the interannual mean and median flow per calendar day for each individual catchment, using data from the calibration period only. We then evaluate the performance of both of these data-based models during the evaluation period and choose the highest KGE score as our benchmark for that particular catchment. This benchmark KGE score gives a sense of how predictable the flow in each catchment is using only streamflow observations at the same temporal resolution as the models use. It represents the minimum accuracy score we expect from any of the conceptual models before considering them as plausible model structures for a given catchment.

2.2. Model Structure Equifinality of Plausible Model Structures

The set of plausible model structures for each catchment contains only those models that beat the daily flow benchmark in that location. These models are different, but all are potentially realistic descriptions of the relevant hydrologic processes in the catchment. However, not all plausible models will beat the benchmark by an equally large margin. We assume that models that outperform the benchmark by a larger margin are better choices to use in a given catchment than models that beat the benchmark by a smaller margin and use this concept to quantify the extent of model structure equifinality. If the best model is joined by multiple other models that exceed the benchmark by a similar amount, model equifinality is considered to be high. Research Question 2 is thus *How many of the 36 model structures in our sample can be considered plausible in a given catchment, and how high is model equifinality within this subset?*

Differences in objective function values are commonly used to quantify model structure uncertainty and equifinality (e.g., Fenicia et al., 2008; Hogue et al., 2006; Winter & Nychka, 2010). We therefore report (1) the KGE score of the best model in each catchment, (2) the difference between the best model's KGE score and the benchmark score in each catchment, (3) the total number of plausible models in each catchment (i.e., the number of models with KGE scores above the benchmark score), and (4) the number of plausible models that fall within 0.01, 0.05, 0.10, and 0.25 KGE value [-] of the best model expressed as cumulative distribution functions (CDFs) across all catchments.

2.3. Relating Differences Between Models to Number of Model Parameters

The number of model parameters is often used to explain differences in model performance. In a study of 19 conceptual models and 429 catchments Perrin et al. (2001) find a tendency for models with more parameters to better fit calibration data but not evaluation data and decreasing robustness (defined as the decrease of mean model performance between calibration and verification periods) for models with more parameters. They suggest that models with more degrees of freedom (i.e., calibration parameters) tend to reproduce errors or noise in the calibration data, a phenomenon called overfitting that is also described by other authors (e.g., Beven, 2012; Schoups et al., 2008; Shaw et al., 2011). Overfitting is related to but not the same as parameter identifiability, which is here used to refer to the ability to identify a unique optimal value for a given calibration parameter. Parameter overfitting leads to an increase in calibration performance combined with a decrease in performance robustness. This is a common expectation when statistical models are used (see, e.g., Figure 12 in Lute & Luce, 2017). This behavior also manifests through evaluation performance that decreases in bias but increases in random scatter for models with a higher number of free parameters (Höge et al., 2018; Lute & Luce, 2017). In conceptual hydrologic models parameters are used as part of equations intended to describe hydrologic behavior, and ideally models that are appropriate descrip-

tions of the dominant processes in a catchment would perform well in such a catchment, regardless of the number of parameters used. Research Question 3 is thus *What are the differences in model performance, and how do these relate to the number of model parameters?*

The analysis for this research question is divided into two parts. First, we investigate the extent to which each model outperforms the benchmark and if any differences between models can be attributed to the number of model parameters. To this end we report (1) the number of catchments in which a model beats the benchmark, (2) the margins by which the model beats the benchmark, (3) how the model ranks compared to the other models in our sample, and (4) the difference in performance between the model and the best model in a given catchment, all in the context of the number of model parameters. A full sensitivity analysis for each combination of model, catchment, and objective function is outside the scope of this work, and we therefore use the total number of calibrated model parameters as a basic surrogate for the effective number of parameters.

Second, we investigate the tendency of models toward overfitting by comparing model performance during calibration, evaluation, and the change between the two periods. We assess this visually with boxplots and use the statistical Mann-Whitney test (Mann & Whitney, 1947) to quantify any differences in a pair-wise comparison of all models. The Mann-Whitney test tests the null hypothesis that two different samples are taken from a single distribution, that is, that $\mu_1 = \mu_2$. If models with more parameters do indeed better fit the data during calibration, we expect that the Mann-Whitney test results show a tendency to reject the null hypothesis for model pairs with a very different number of parameters, combined with a tendency to not reject the null hypothesis for model pairs with a similar number of parameters. The expectations that models with more parameters have lower robustness is tested in the same manner.

The KGE can be decomposed into its three constituent parts, which reflect the similarity between simulations and observations in terms of the correlation between the two, the ratio of standard deviations, and the ratio of means (Gupta et al., 2009). The latter two components can be seen as indications of the scatter and bias of the simulations and investigate these as a second test for overfitting, both visually as box plots and through Mann-Whitney tests (Mann & Whitney, 1947). Our expectations for the Mann-Whitney test are as outlined in the previous paragraph.

2.4. Model Suitability for Different Catchments

Model development generally takes places on geographically small scales, such as one or a few research catchments where in-depth knowledge of catchment conditions can inform the choice of model structure (e.g., Ambroise et al., 1996; Fenicia et al., 2016; McGlynn et al., 2002; Peters et al., 2003). Across larger scales, the relation between climatic conditions and conceptual model performance has been well studied (e.g., Dakhlaoui et al., 2017; Fowler et al., 2018; Lidén & Harlin, 2000; Merz et al., 2011; Van Werkhoven et al., 2008; Van Esse et al., 2013). Catchment-averaged attributes beyond climatic data have proven useful to to assess conceptual model strengths and weaknesses across the United Kingdom (Lane et al., 2019). Such studies are generally conducted with a limited number of models, a limited number of catchments, or within a geographically small and thus relatively similar area (such as Austria, Merz et al., 2011, and France, Van Esse et al., 2013). Until recently no data were available to allow such studies to also investigate the relationship with catchment structure across a large and varied domain. The CAMELS data set provides catchment attributes spread across six main categories: climate, geology, topography, soil, land cover, and streamflow (Addor et al., 2017). We attempt to use these descriptors to clarify the relation between model performance and catchment type. Research Question 4 is thus *How does relative model performance relate to known catchment attributes?*

Because efficiency scores are not easily compared between places (e.g., Schaefli & Gupta, 2007), we instead rank the plausible model structures in each catchment based on their KGE scores and use model ranks for this analysis. Given the size of the data sample, we limit this aspect of the study to an exploratory analysis based on Spearman rank correlations between model ranks and catchment attributes only.

3. Data and Models

3.1. CAMELS Catchment Data

This study uses the CAMELS data set (Addor et al., 2017), which provides time series of meteorological variables and streamflow (Newman et al., 2015), and tables with catchment attributes for 671 catchments

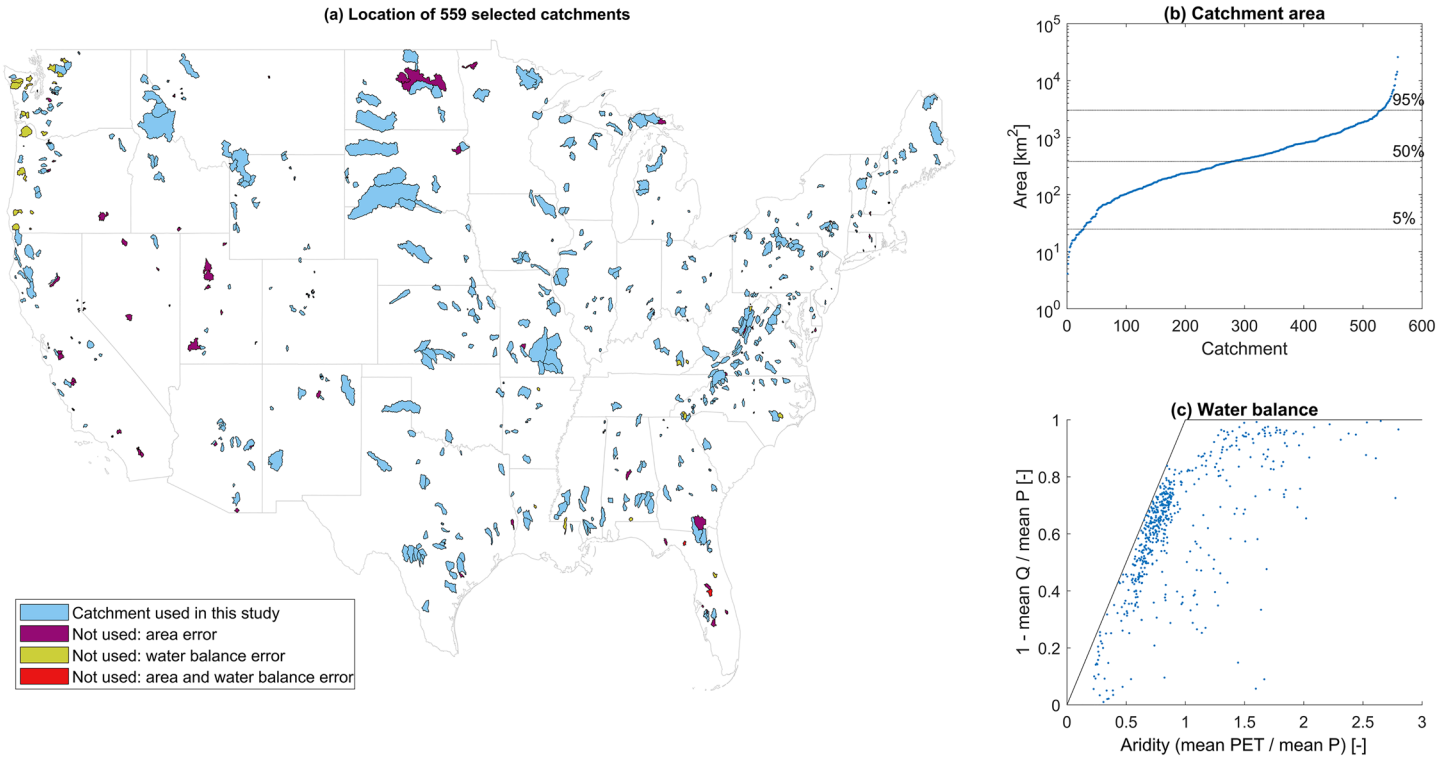

Figure 1. Catchments in the CAMELS data set. Five hundred fifty-nine catchments were used in this study (blue), after removing those catchments with uncertain area estimates (purple; >10% differences between two geospatial data sets and the USGS reported value) or water balance errors (yellow) or both (red). (a) Geographical location and reason for exclusion from this study. (b) Catchment area distribution of the 559 selected catchments. (c) Aridity index against 1 − runoff ratio for the 559 selected catchments.

in the contiguous United States. We perform several basic data checks and remove those catchments with large (>10%) discrepancies between catchment area as used for averaging of the meteorological time series and area as published by the USGS (US Geological Survey, 2018) or the higher resolution GAGES II data set (provided as part of the CAMELS data set). We use preliminary screening (e.g., Martinez & Gupta, 2011) to remove those catchments that fall outside the energy limit and water limit on the Budyko curve (Budyko, 1974). This leaves 559 catchments for use in this study, distributed across the contiguous United States (Figure 1).

The CAMELS data provide three different forcing products (Newman et al., 2015) at a daily resolution. This study investigates lumped models (catchments are treated as a single entity) and thus uses catchment-averaged forcing data. We follow Newman et al. (2015) and Addor et al. (2017) in using the Daymet product, which is based on the highest spatial resolution of all three products (1 km × 1 km compared to 12 km × 12 km for Maurer and NLDAS products) and is more likely to provide accurate data for smaller catchments and locations with complex topography. Time series of daily precipitation and temperature are part of the CAMELS data, and time series of potential evapotransiration (PET) are estimated using the Priestley-Taylor method (details in supporting information Text S1 Priestley & Taylor, 1972).

3.2. MARRMoT Modeling Framework

This study uses 36 conceptual model structures from the Modular Assessment of Rainfall-Runoff Models Toolbox (MARRMoT) v1.0 (Knoben et al., 2018a, 2019), which organizes conceptual model code in a single uniform framework. This has the main advantage that the implementation of models and fluxes is consistent and any differences in simulation are thus solely due to differences in model structure. The MARRMoT models used in this work are all based on published literature and cover a wide range of possible structures, from a simple one-parameter model to structures with up to 6 stores or 15 parameters. The toolbox is provided with literature-based parameter ranges for each model to support parameter sampling or optimization. These standardize the parameter ranges as much as possible, so that models have the same amount of parameter freedom (e.g., in the case of interception capacity, all models that simulate the interception process use a range of 0–5 mm). The differential equations that express each model's changes

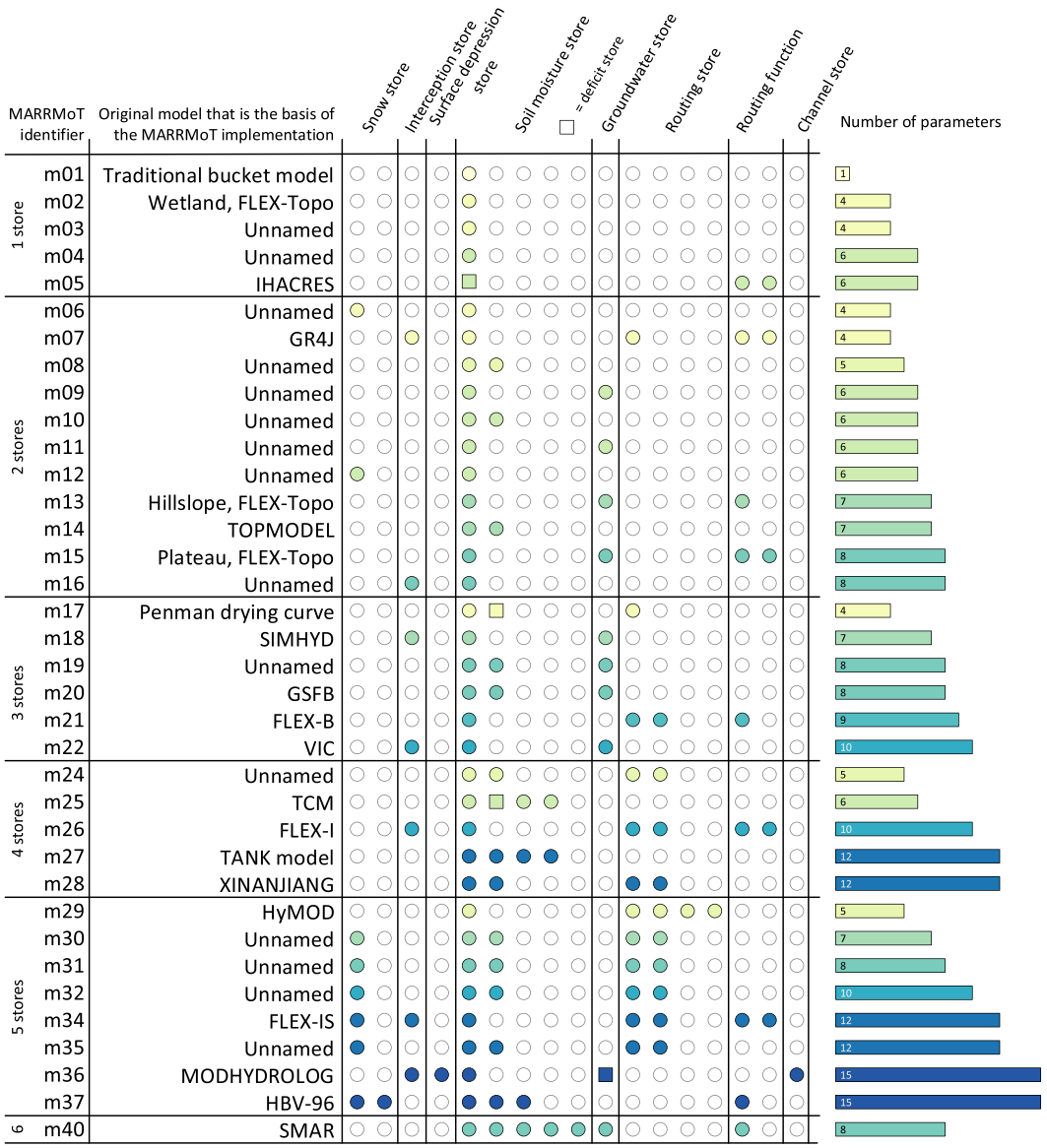

Figure 2. Summary of the 36 models used in this work (adapted from Knoben, Freer, Fowler, et al., 2019, Figure 2). Models are sorted by the number of stores first (indicated in the left column) and by their number of parameters second (bars in right column). MARRMoT identifier refers to an identifier that is used in the MARRMoT documentation and subsequent analysis in this work. Identifier, number of parameters (p) and number of stores (s) are used in other figures as, for example, "m01 (1p, 1s)." The middle part of the figure gives an overview of the processes each model's store(s) is intended to represent. Note that MARRMoT model implementations can deviate from the source models they are based on. See the MARRMoT documentation for details.

in storage(s) with time are numerically approximated with a fixed-step implicit Euler method, which uses the same step size as the forcing data (detailed settings for reproducibility can be found as part of the data package that accompanies this paper). The implicit Euler method provides better accuracy and stability compared to the Explicit Euler method, at the cost of increased computational times (Kavetski et al., 2006; Schoups et al., 2010). Figure 2 provides an overview of the 36 models used in this work.

3.3. Model Setup

Data for each catchment are divided into two 10 year periods covering 1 January 1989 to 31 December 1998 (calibration) and 1 January 1999 to 31 December 2009 (evaluation), respectively. Average climate characteristics are approximately constant between these periods with the exception of regions with high mean precipitation ($\bar{P} \geq 5$ mm/day; precipitation has decreased somewhat) and regions with low mean

temperatures ($\bar{T} \leq 5°C$; temperatures have increased). Estimated potential evapotranspiration rates are approximately constant between the two periods (see Figure S1). Streamflow records are complete during this period for 546 catchments. For the remaining 13 catchments, days with missing streamflow values are ignored during the calculation of objective function values. Missing values account for 0.013% to 4.8% of all observations in these catchments.

The 36 models in this study are calibrated for 559 catchments using three different objective functions, each based on the KGE (Gupta et al., 2009):

$$\mathrm{KGE} = 1 - \sqrt{(r-1)^2 + \left(\frac{\sigma_{sim}}{\sigma_{obs}} - 1\right)^2 + \left(\frac{\mu_{sim}}{\mu_{obs}} - 1\right)^2} \quad (1)$$

where subscripts *obs* and *sim* refer to observed and simulated time series of flow, respectively, r is the linear correlation coefficient between observed and simulated flow, σ denotes the standard deviation of flows, and μ the mean of flows. We aim to compare model performance for a variety of flow conditions; therefore, our choice of objective functions emphasizes higher flows, lower flows, and a combination of both. The objective functions used are the KGE calculated on time series of flow (KGE(Q)), the KGE of inverse flows (KGE($1/Q$), and the mean of KGE(Q) and KGE($1/Q$). Inverse flows are shown to be more appropriate than a log transform to emphasize low flows (e.g., Pushpalatha et al., 2012; Santos et al., 2018). Pushpalatha et al. (2012) add a constant e to observed and simulated streamflow to avoid problems with inverting zero flow values. They recommend e to be set at 1% of the mean observed flow, because this limits the impact of the added constant on the resulting NSE($1/Q$) values in their study. Because no such guidance yet exists for KGE and because NSE and KGE are conceptually based on the same three components (Gupta et al., 2009), we assume that this value is an appropriate choice for KGE($1/Q$), too.

We use the Covariance Matrix Adaptation Evolution Strategy (CMA-ES Hansen, 2016; Hansen & Ostermeier, 1996, 2001; Hansen et al., 2003) to calibrate model parameters. CMA-ES is a single-objective optimizer that compares favorably to various other methods for finding the global optimum of difficult functions and in rugged objective function landscapes (Arsenault et al., 2014; Hansen et al., 2003, 2010). The algorithm has seen successful application in hydrology (e.g., Arsenault et al., 2014; Fowler et al., 2018; Peterson & Western, 2014), as well as many other fields (Hansen, 2009). The algorithm is allowed to run either until the change in objective function of all members in the current generation and the range of objective function values in at least the preceding 10 generations is below 1E−3 or until the standard deviation of the normal distribution used to sample parameter values for the new generation drops below 1E−3. These are problem-dependent algorithm settings (Hansen, 2016) that we consider an acceptable compromise between accuracy and speed for the 60,372 combinations of models, catchments, and objective functions. Details about CMA-ES stopping criteria and algorithm exit flags can be found in supporting information Text S2.

Model warm-up periods are used to reduce the impact of uncertain initial conditions on model performance. Recent studies have attempted to provide guidelines for warm-up period length in conceptual models (Kim et al., 2018), but these studies are limited in number of models (1 and 2, respectively) and catchments (18 and 1, respectively), and it is therefore difficult to generalize their findings to a large-sample study such as this. Instead of using a fixed number of warm-up days, we determine the initial storages in an iterative procedure by letting the model repeat Year 1 of the data period until the stores reach an equilibrium for the first day of the year (<1% change in storage value[s] between runs). Storage values might not converge for certain parameter sets (e.g., when a store of unlimited depth has very low outflow), in which case the procedure is stopped after 50 iterations.

4. Results

Results presented here are based on data for the KGE(Q) objective function obtained from the evaluation period, unless specifically indicated as being calibration results or relating to one of the two other objective functions.

Where applicable, findings for each model only include results from those catchments where the model exceeds the minimum benchmark level of expected performance. Each section also includes a brief summary of findings for the other two objective functions, KGE($1/Q$) and 1/2*[KGE(Q) + KGE($1/Q$)]. Figures for these objective functions are part of the supporting information for brevity.

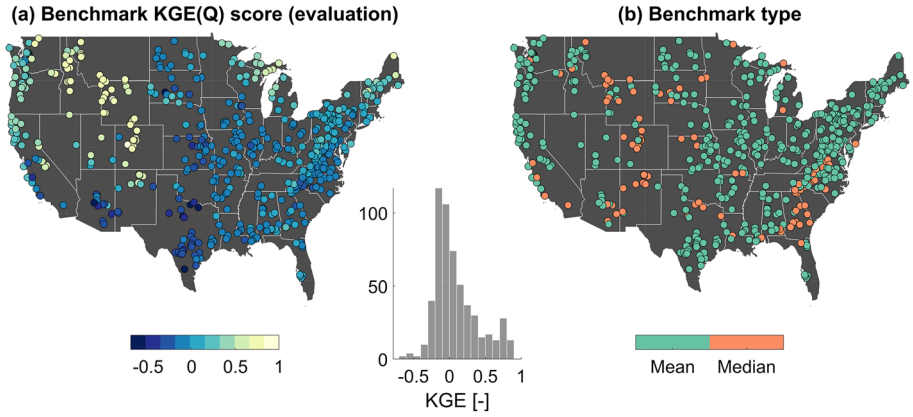

Figure 3. (a) Benchmark KGE(Q) score that a model must beat to be considered a plausible model structure in each catchment. The benchmark is treated as any other model in the sense that the benchmark simulations are calculated from flow observations in the calibration period, and the benchmark KGE(Q) score is calculated by comparing these simulations to observations from the evaluation period. (b) Type of benchmark (mean or median calendar day flow) that gives the higher benchmark KGE(Q) score.

4.1. Defining a Lower Level of Expected Model Performance

The performance of the benchmark time series (i.e., the mean or median daily flow regime) varies across space and is subject to strong spatial organization (Figure 3a). KGE scores are lowest (KGE ≤ -0.5) in very arid areas and highest in snow-dominated areas (with values up to KGE = 0.89). Approximately 80% of these benchmarks are obtained by using the mean calendar day flow, the remainder being obtained from the median calendar day flow (Figure 3b). These results set a baseline for minimum expected model performance by indicating how predictable the flow regime is.

This spatial organization is also visible for the KGE($1/Q$) and 1/2*[KGE(Q) + KGE($1/Q$)] objective functions (Figures S3 and S4). Benchmark values for KGE($1/Q$) are generally higher than those for KGE(Q) and are in 96% of cases obtained by using the median calendar day flow. As expected, the results for the $\frac{1}{2}$[KGE(Q) + KGE($1/Q$)] objective function are in between the results of the other two.

4.2. Model Structure Equifinality of Plausible Model Structures

Figure 4a shows that the maximum achieved evaluation efficiency in each catchment (i.e., what the best model out of 36 achieves) is subject to strong spatial organization, although exceptions to the pattern exist. Maximum efficiency ranges from −0.11 to 0.93. Note that these values are raw KGE scores and are not yet adjusted by the benchmark score. In geographical terms, maximum model performance tends to be lowest in the central United States (plains areas east of the Rocky Mountains) and certain parts of the southwest. These areas share a tendency to be very arid (see Figure 3c in Addor et al., 2017). Figure 4b shows the number of models that fall within certain performance thresholds. Curves that stay closer to the bottom indicate that fewer models have a KGE value within 0.01/0.05/0.10/0.25 of the best model in a given catchment. For example, the blue line indicates that in approximately 350 catchments, no model has a KGE value within 0.01 of the best model, while in the remaining 200 catchments at least one and up to eight models have performance within 0.01 of the best model for each catchment. These results show that model structure equifinality can be very high: In many catchments several models can be virtually indistinguishable (within 0.01 KGE of each other) in terms of efficiency scores, and in the vast majority of catchments up to 28 different models can be close (within 0.05 KGE) to the best model.

Comparing maximum model efficiency to the predefined benchmark values in each catchment provides context for the maximum model efficiency scores (compare Figures 4a and 4c). The difference between maximum model performance and benchmark is smallest in mountainous regions and generally high in arid regions. In 11 catchments no model outperforms the benchmark ($\Delta KGE < 0$), all characterized by a high fraction of precipitation occurring as snowfall. In these places, no model in our sample is able to simulate the persistent features of the hydrograph (i.e., features that recur every year) better than the benchmark does.

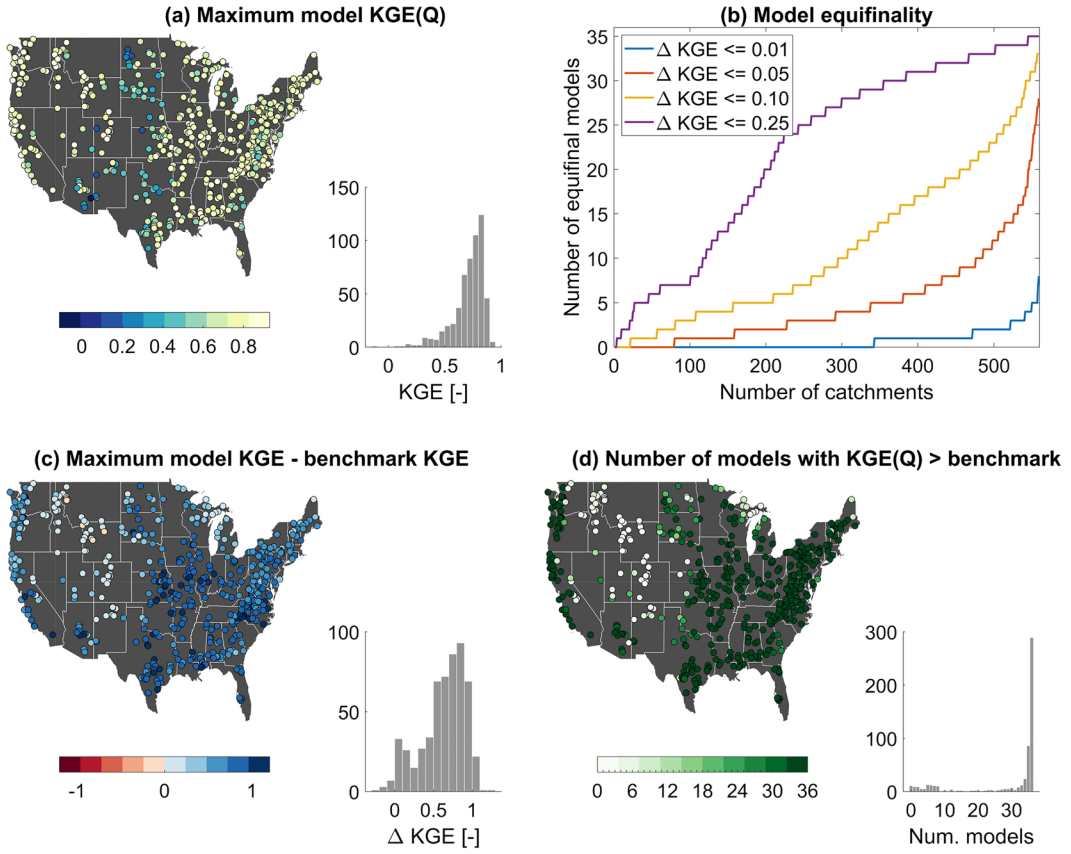

Figure 4. Results shown for the evaluation period. (a) Maximum model efficiency, showing the highest KGE score obtained in each catchment (note that this is not necessarily by the same model). (b) Model equifinality, showing how many models are within 0.01, 0.05, 0.10, and 0.25 KGE distance away from the best model in each catchment. Note that the catchments are sorted independently for each CDF and that lines should not be compared on a per-catchment basis. (c) Difference between maximum model efficiency and benchmark efficiency in each catchment. (d) Number of models that beat the benchmark in each catchment.

These 11 catchments are excluded from further analysis, because our model ensemble contains no structures that meet our plausibility criterion. Figure 4d shows that in mountainous regions the number of models that beat the benchmark is low. This can partly be explained by not all models having a snow module (only eight models do), but even having a snow module is no guarantee that a model can beat the benchmark. In contrast, in wet, nonsnowy regions the vast majority of models beats the benchmark and in 289 out of 559 catchments every single model in our sample provides more accurate simulations than the benchmark gives (although that does not automatically imply that all model simulations are equally close to observations in these catchments, see Figure 4b). Model choices matters most in the arid regions where maximum model efficiency is lowest. Models do exist that can provide reasonable simulations here, but they must be carefully selected.

Spatial patterns of maximum model KGE and KGE distributions are roughly similar for the three objective functions (see Figures S5 and S6). Maximum evaluation efficiency ranges from −0.74 to 0.96 for the low flow objective function ($KGE(1/Q)$) and −0.15 to 0.91 for the combined flow objective function ($\frac{1}{2}[\mathrm{KGE}(Q) + \mathrm{KGE}(1/Q)]$). Model equifinality is lower (i.e., fewer models are close to the best model in each catchment), especially for the combined objective function. There are more catchments (19 and 24, respectively) where no model beats the benchmark and fewer catchments where most models can beat the benchmark (in only 189 and 200 catchments, respectively, do more than 30 models beat the benchmark). Many models struggle to achieve accurate low flow simulations but the maximum model KGE scores for the $KGE(1/Q)$ objective show that this is not impossible, only that it requires careful model selection. Adopting a multiobjective approach reduces model equifinality by the largest degree.

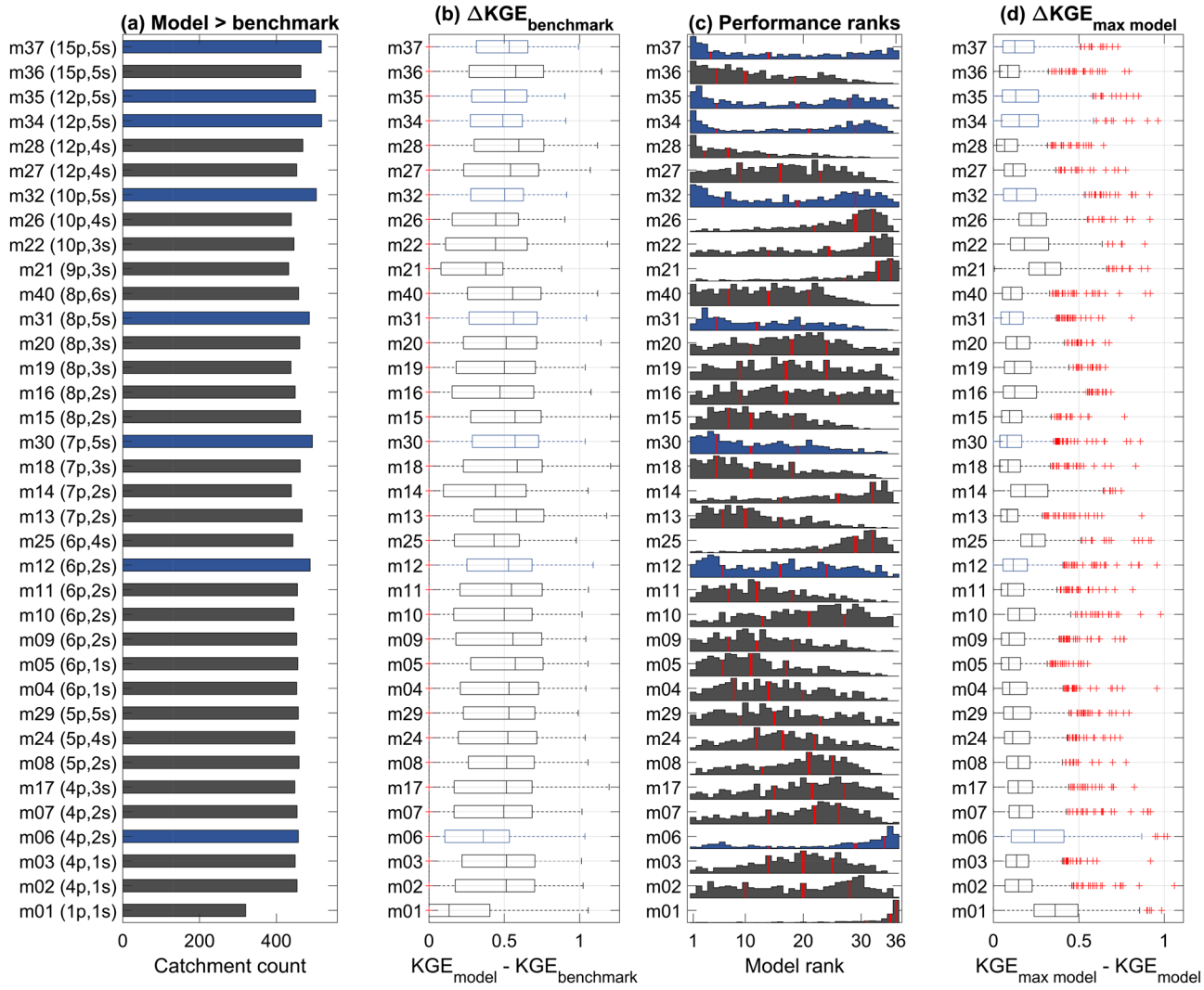

Figure 5. Results shown for the evaluation period. (a) Number of times each model beats the KGE(Q) benchmark. Models with a snow module are shown in dark blue. Models are sorted by number of parameters first and number of state variables (stores) second. For example, m37 (15p, 5s) means that the model with identifier m37 has 15 parameters and 5 stores. (b) Box plot of the margin by which each model beats the benchmark. Negative values (cases where the model does not beat the benchmark) are not shown. (c) Histograms of model ranks, where Rank 1 indicates the model with the highest KGE evaluation score and Rank 36 indicates the model with the lowest KGE evaluation score. Note that histograms are scaled individually to best make use of the available space and that bar heights should not be compared between rows. Red lines indicate 25th, 50th, and 75th percentiles. (d) Box plot of the difference between the model's performance and that of the best model in each catchment.

4.3. Relating Differences Between Models to Number of Model Parameters

4.3.1. Differences Between Models During Evaluation

Figure 5 compares performance of individual models during evaluation. With the exception of the simplest Model m01, models beat the benchmark in approximately equal numbers (Figure 5a). Models that include a snow module (shown as blue bars) tend to beat the benchmark in more catchments than models without a snow module for obvious reasons. There is substantial variety in the margin by which models beat the benchmark (Figure 5b), showing that certain models are much better suited to flow simulation with the KGE(Q) objective function than other models are. This is reflected in the ranks these models obtain (Figure 5c). Number of parameters is a poor predictor of how a model will perform. Certain models, such as m36 (15 parameters), m28 (12 parameters), and m13 (7 parameters), tend to rank better (toward Rank 1), whereas other models, such as m26 (10 parameters), m21 (9 parameters), m25 (6 parameters), and m06 (4 parameters), tend to rank much worse (toward Rank 36). Many models with a snow module show bimodal distributions, indicating that in certain catchments (i.e., those with snow) they are one of the best options available, but this does not imply they are among the better choices in other catchments. Differences in

model suitability for the KGE(Q) objective function can also be seen in how far away each model tends to be from the best model in each catchment (Figure 5d). Models such as m28 (12 parameters), m13 (7 parameters), and m05 (6 parameters) tend to perform similar to the best model in any catchment, whereas models such as m34 (12 parameters), m21 (9 parameters), and m06 (4 parameters) tend to be much further from the best model in each catchment. These results suggest that the choice of model parametrization (i.e., which equations are used and how parameters are used within them) is more important to dictate a model's strengths and weakness than how many parameters the model has.

There is greater variety in the number of catchments in which a model beats the benchmark for the other two objective functions (see Figures S7 and S8) and broadly speaking these results support the conclusion that certain model structures are much better suited for certain objective functions. Of particular note are Models m28, m27, and m21 because they showcase three different model types: m28 performs well under all three objective functions; m27 performs reasonably well with the KGE(Q) objective function but performs substantially worse for the other two objectives; m21 performs poorly on the KGE(Q) objective function but performs very well on the KGE($1/Q$) objective, moving from being consistently one of the worst choices to consistently one of the best.

4.3.2. Tests for Overfitting

Contrary to expectations, a higher number of model parameters does not necessarily lead to higher efficiency values during calibration (Figure 6a). In other words, models with higher degrees of freedom (more parameters) are not consistently better at fitting the calibration data. In fact, several of the models that show lower calibration efficiency ranges (e.g., m21 and m26) have a relatively high number of free parameters (9 and 10, respectively). Both simpler (e.g., m02, four parameters) and more complex models (e.g., m35, 15 parameters) show higher ranges of efficiency values during calibration.

Evaluation performance shows a similar pattern (Figure 6b): There are certainly differences between the ranges of efficiency values obtained by different models, but this seems unrelated to the number of parameters each model has. Overall, evaluation efficiency ranges are somewhat lower than calibration ranges, which indicates either a change in catchment conditions that the models insufficiently account for (e.g., change in climatic forcing), or a degree of overcalibration (i.e., the models are calibrated to a certain amount of data noise). Analysis of each model's performance change between calibration and evaluation periods (Figure 6c) shows that, whatever the cause, distributions of performance change are similar across all models. Figure 6c also shows that performance decline during evaluation does not always occur, and in approximately a quarter of all catchments model performance instead increases during evaluation (note that these are not necessarily the same catchments for each model). While this may seem like a high number, we note that many studies conducting similar analyses deliberately choose periods with contrasting climate between calibration and evaluation periods and find declining model performance under contrasting conditions, whereas here climatic conditions are relatively similar in both periods (see Figure S1).

Pair-wise Mann-Whitney statistical tests (see Figure S9) confirm that there are certainly differences between the distributions of model performance but that there are no clear patterns that relate to the number of model parameters. For example, the null hypothesis that calibration performance of Models m17 (4 parameters) and m37 (15 parameters) are drawn from the same distribution cannot be rejected ($p > 0.95$). Analysis on a per-catchment basis (not shown for brevity) also shows that no significant ($p < 0.05$) relation exists between either calibration performance, evaluation performance or robustness, and the number of model parameters.

Analysis of the constitutive KGE components (correlation, scatter, and bias; Figure S10) shows that the expectation that bias decreases while scatter increases for models with a higher number of parameters (see, e.g., Höge et al., 2018) is not a general rule that can be applied to these conceptual models. Collectively, models show a tendency to overestimate the bias component ($\mu_{sim} > \mu_{obs}$, although exceptions such as m01, m17, m14, and m22 exist; Figure S10c). Models also show a tendency to overestimate the scatter component ($\sigma_{sim} > \sigma_{obs}$; Figure S10b) but again exceptions exist. The clearest variability can be seen in the correlation component (Figure S10a) where certain models score substantially lower than others, but here too no relation with number of parameters can be seen.

Pair-wise Mann-Whitney tests (see Figure S11) indicate that distributions of values for KGE components can be different for different models but that results for models with more parameters are not necessarily

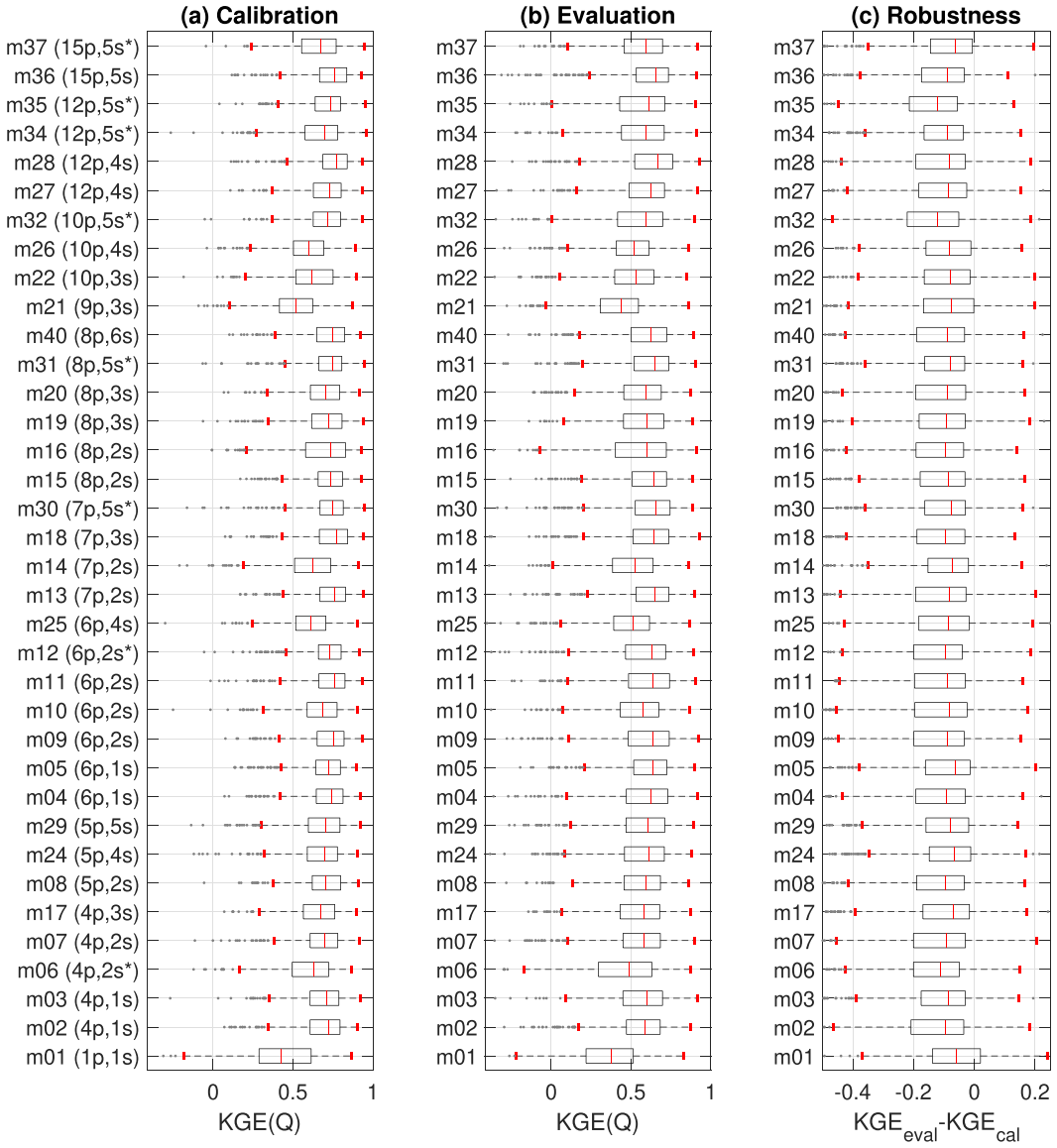

Figure 6. (a) Model performance during calibration. (b) Model performance during evaluation. (c) Model performance robustness defined as the change in performance between calibration and evaluation. Results are only shown for catchments where each model beats the benchmark during evaluation. Models are sorted by number of parameters first and number of state variables (stores) second. Models that include a snow module are indicated with an asterisk (*).

statistically different from results from models with fewer parameters in a consistent way. These results support the findings in the previous section and again suggest that the number of calibrated model parameters is not an effective measure for explaining differences in conceptual model performance.

The main conclusions can be found for the other two objective functions as well: Models perform quite differently, but this cannot be consistently explained by an increasing number of parameters (see Figures S12–S19). Interestingly, plots of the KGE components for both other objective functions (see Figures S16 and S18) show that most models have a tendency to underestimate the bias and scatter components (i.e., the opposite of what happens with the KGE(Q) objective), showing the impact of objective function choice on model simulation errors.

4.4. Relating Model Performance to Knowledge About Catchments

4.4.1. Correlations Between Model Ranks and CAMELS Catchment Features

The 52 CAMELS catchment features are divided into six categories: climatic conditions, observed streamflow signatures, geologic properties, topographic properties, vegetation properties, and soil properties. The

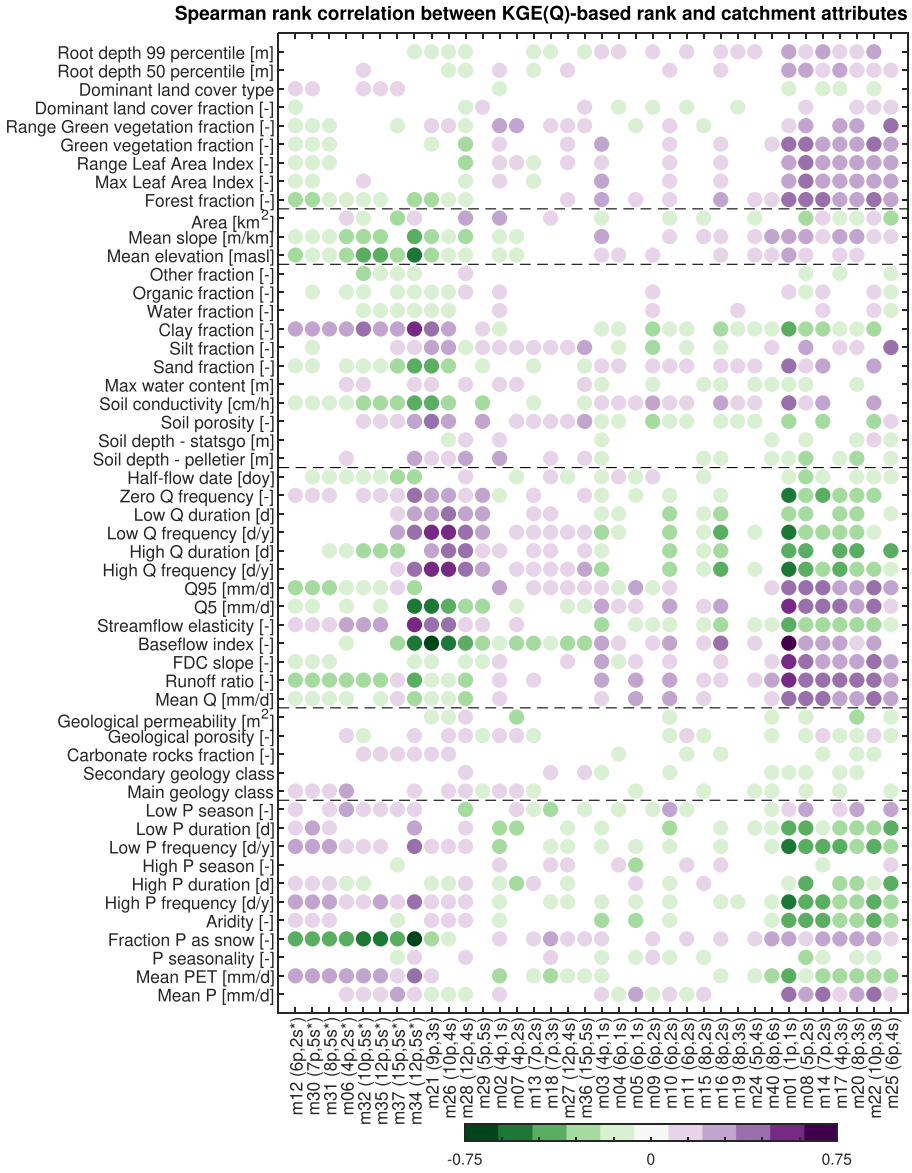

Figure 7. Spearman rank correlation between model ranks in the evaluation period and the different categories of CAMELS catchment data (separated by dotted lines, these are properties of vegetation, topography, soil, streamflow, geology, and climate). For each model, correlations are only calculated for catchments where the model beats the benchmark. Models are ranked from best to worst, with Rank 1 indicating the best model. Only correlations with p value <0.05 are shown, and color intensity corresponds to the strength of the correlation. Models are sorted manually in an attempt to place models with similar correlation patterns close together. Example interpretation: See the dark green points for the combination of models with a snow module (indicated with an asterisk (*)) and "fraction P(recipitation) as snow" that indicates a strong negative correlation (bottom left of the figure). As the observed snowfall fraction of catchments increases, models with a snow module tend to achieve lower rank numbers, that is, toward Rank 1, and thus rank better than the other models in our sample.

categories are not fully independent. The connection between climatic conditions and streamflow regimes on continental to global scales is well established (e.g., Addor et al., 2018; Berghuijs et al., 2014; Knoben, Woods, and Freer 2018; Kuentz et al., 2017), and this shows in the CAMELS data as high correlations between climatic conditions and observed streamflow signatures (Addor et al., 2018, see also Figure S20 for cross correlations). Climatic conditions and to a lesser extent observed streamflow signatures also correlate strongly with vegetation attributes. Geologic and soil properties contain the most independent information. Scatter plots of model ranks and CAMELS catchment attributes (not shown for brevity) indicate that empirical relationships exist between model ranks and certain types of catchments but also that substantial

scatter around any main relationship is present. Figure 7 summarizes these relationships using the Spearman rank correlation coefficient. Note that for each model only those catchments are included where the model beats the benchmark.

The strongest correlations for most models can be found with observed streamflow signatures (for a description of these signatures, see Addor et al., 2018) and to a slightly lesser extent with climatic conditions. If streamflow signatures are seen as a way to describe flow regimes, this suggest that certain models are relatively more or less suited for certain flow regimes (compared to the other models in our study). The established connection between streamflow regimes and climatic conditions explains why model ranks also correlate with climatic conditions. Correlations with those CAMELS attributes that describe the catchments' geology, topography, soil, and vegetation are generally the weakest, except in cases where those attributes also correlate with climatic conditions or observed streamflow signatures. For example, mean elevation (topography) correlates strongly with fraction precipitation as snowfall (climate). This can imply several things: (1) Uncertainty in the attributes data is too high to find any clear relationships with model performance; (2) we are not looking at the right catchment attributes because these do not seem to explain the hydrologic and model behavior; (3) models work better/worse for certain streamflow regimes, but regimes are not a unique result from a certain arrangement of catchment attributes.

These conclusions are similar for the other two objective functions. Correlations for the KGE($1/Q$) objective function (see Figure S21) are generally lower than those in Figure 7, but the strongest correlations can be seen between model ranks and observed streamflow signatures. Correlations for the $\frac{1}{2}$[KGE(Q)+KGE($1/Q$)] objective function (see Figure S22) are similar to those in Figure 7, in terms of both pattern and strength.

4.4.2. Model Structure Similarity

In Figure 7, models are sorted manually along the x axis, such that models with similar correlation patterns in the y direction are placed close to one another. This allows us to define model groups that contain model structures with similar performance ranks across the sample of catchments.

An obvious relation exists between models that include a snow component (m12 to m34, leftmost on the x axis) and catchments where a larger fraction of the annual precipitation occurs as snowfall, where these models naturally achieve higher efficiency scores than models without the capability to simulate snow accumulation and melt. The next group consists of Models m34, m21, m26, m28, and m29. These perform relatively better in baseflow-dominated catchments without flashy streamflow behavior. These particular models share a structural feature that consists of a soil moisture routine that simulates a variable contributing area, which then drains into a linear reservoir. Models m01, m08, m14, m17, m20, m22, and m25 (at the right on the x axis) share a tendency to rank better in catchments with low precipitation, low mean flows and low flows (Q5) and a larger number of high precipitation events. This suggests drier catchments with low flows punctuated by the occasional high flow event. These models all have a mechanism that allows incoming precipitation to reach the stream quickly (either saturation excess or a bypass mechanism) and also contain a mechanism that ensures that very low (up to 0) flows can be generated. This mechanism is either threshold-based, where no flow is generated unless a storage threshold is exceeded, or evaporation-based, where evaporation can occur from multiple stores and can thus be used to prevent water from reaching the stream. Model m01 is the most extreme member of this group, containing only a saturation excess mechanism. The models in this group suggest that very different model structures are capable of reproducing similar flow regimes. The remaining models can be roughly divided into two groups: Models m02 to m36 and m03 to m40. These models do not share any obvious characteristics and do not show many pronounced correlations.

Similarity of correlation patterns can be seen for the other two objective functions too (see Figures S21 and S22) but model groupings are different. This might imply that different parts of the model structure influence whether model structures behave similarly on a given objective function. Further analysis for the other two objective functions is considered out of scope for this work.

5. Discussion

5.1. Synthesis

Large-sample analysis such as this study can provide unique insights into our ability to model a wide variety of catchments and how models differ from one another in a practical sense. Here, we return to the research questions posed in section 2.

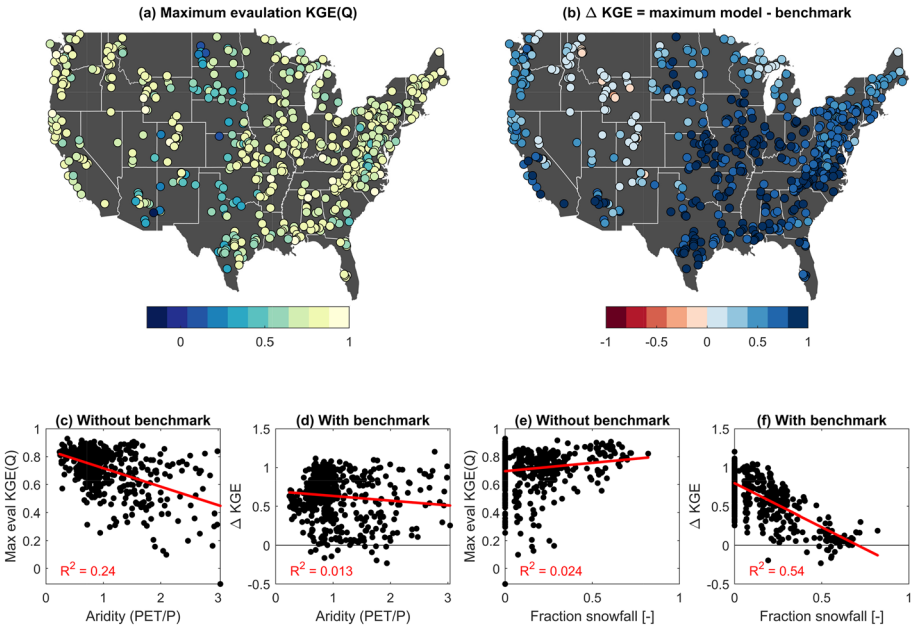

Figure 8. (a) Maximum model performance during evaluation. (b) Difference between maximum model performance and benchmark KGE. (c–f) Relations between model performance and climatic conditions, showing how (not) using a benchmark can affect conclusions.

We first answer the question "What level of model performance can reasonably be expected in a given catchment?" by defining benchmarks based on daily flow observations in the calibration period (Figure 3). This results in benchmark evaluation KGE scores that range between −0.65 and 0.87, with 98.8% of catchments having benchmark values that are higher than what would be obtained by using the mean annual flow (i.e., KGE $= 1-\sqrt{2}$; Knoben, Freer, & Woods, 2019). The benchmark scores gives us the necessary context to evaluate model performance, by showing what the typical seasonal signal in the data is and which efficiency scores can easily be obtained in a given catchment. Benchmark scores show strong spatial organization and are typically highest in snow-dominated catchments and lowest in arid catchments.

Answers to our second question, "How many of the 36 model structures in our sample can be considered plausible in a given catchment, and how high is model equifinality within this subset?" are dependent on where we draw the line between plausible and implausible models. Without an explicit statement about the benchmark that we expect our models to beat we might have concluded that our model sample performs worst in arid catchments (see the strong negative gradient in Figure 8c). In fact, this would have been in line with existing literature that states that it is harder to obtain high efficiency scores in arid locations than it is to obtain such scores in more humid regions (e.g., Fowler et al., 2018; Krysanova et al., 2017; Melsen et al., 2018; Newman et al., 2017; Van Esse et al., 2013). With our specified benchmark arid regions do not stand out as places where our model ensemble does poorly. Despite the challenges to hydrologic modeling in arid regions (Pilgrim et al., 1988), our model ensemble is able to beat the benchmark in arid and humid regions by approximately equal margins (Figure 8d). Instead, we have reason to doubt the ability of the models in our sample to simulate cold-region hydrology: Models tend to achieve higher KGE evaluation scores as fraction snowfall increases but improvement over benchmark drastically lowers. Of course, when the benchmark scores increases there is less potential for improvement to be achieved by any model and lower ΔKGE values might be expected (e.g., with a benchmark KGE = 0.95, the potential for improvement is only 1−0.95 = 0.05). However, in 11 snow-dominated catchments our model ensemble is unable to reproduce the persistent seasonal streamflow signal identified by the benchmark model and no model in the ensemble can beat the benchmark score. This might suggest any combination of missing or inappropriate process representations, issues with the input data or problems with model calibration. In the vast majority of nonsnowy catchments many to all of the models can beat the benchmark (Figure 4). They do not do so by equal margins but in approximately 200 catchments up to 8 model structures achieve practically the same efficiency score as the best model (<0.01 KGE difference) and in approximately 500 catchments up to 28 models can be close (<0.05 KGE difference) to the best model. Logically, in cases where model structure equifinality is high, not

every model can be an equally good representation of the catchment under consideration and that these models produce hydrograph simulations of similar accuracy does not mean that they do so for the right reasons (Kirchner, 2006). These results provide evidence from a very large sample of model structures and catchments that better assessment of model structural adequacy (e.g., Gupta et al., 2012) and process fidelity in models (e.g., Clark et al., 2016; Kirchner, 2006) should become the norm. Relying on aggregated efficiency scores alone is insufficient to determine which models are appropriate choices for which catchments.

We next answer the question "What are the differences between models, and how do these relate to the number of model parameters?" The expectation that models with more parameters are vulnerable to overfitting cannot be seen in our results. This contrasts with findings by Perrin et al. (2001), who reported a tendenc for conceptual hydrology models with more parameters to better fit calibration data but not evaluation data, and an inverse relation between model robustness (defined in their paper as the decrease of mean model performance between calibration and verification periods) and number of model parameters. This pattern is not visible in our sample (Figures 6, S12, and S14) and not found by our use of statistical tests (Figures S9, S13, and S15). Equally, the expectation that models with more degrees of freedom generate errors with reduced bias and increased scatter (Höge et al., 2018; Lute & Luce, 2017) is not seen in the constitutive components of the KGE objective function (Figures S10, S11, and S16–S19). While overfitting (i.e., performance loss in evaluation due to noise fitting during calibration, e.g., Beven, 2012; Schoups et al., 2008; Shaw et al., 2011) is a clear issue with high-degree polynomials (Grayson & Blöschl, 2001; Schoups et al., 2008), these principles do not seem to apply to our sample of conceptual hydrologic models and catchments. Consequently, the number of parameters might only be a good way to quantify model complexity in the restricted case that the models with more parameters contain the models with fewer parameters as a special case (e.g., a fifth-order polynomial contains fourth-order polynomials as a special case). This condition is generally not met in model intercomparison studies which typically aim for diversity, not similarity, in the models included (e.g., see the models used in Franchini & Pacciani, 1991; Perrin et al., 2001; Seiller et al., 2012; and this study).

Whereas most models show a tendency to perform well on specific objective functions but not on others (a common finding in studies comparing multiple models across multiple objectives; see, e.g., Fowler et al., 2018; Perrin et al., 2001; Seiller et al., 2012), certain models seem to display more well-rounded behavior and tend to rank better regardless of the objective function used. It is therefore noteworthy and somewhat unexpected to find that Model m28 is consistently among the best, if not the best, model structure in the majority of catchments and for all three objective functions. This model is the MARRMoT version of the Xinanjiang model (Knoben et al., 2019; Zhao, 1992), modified with a unique feature not seen in any other model in our sample, namely a double parabolic curve that is used to represent the fraction of the catchment that contributes to free drainage (Jayawardena & Zhou, 2000). Nonlinear treatment of saturated area representation has been linked to more flexible model performance within groundwater-dominated catchments before (Lane et al., 2019) and we can speculate that this specific double parabolic formulation gives the model a unique capability that allows it to perform well in a wide variety of catchments. Interestingly, it is difficult to generalize these findings because for every model (including m28) certain catchments can be found where that model is one of the best structures (in terms of efficiency scores during evaluation) and equally catchments can be found where that model is one of the worst options (Figure 9 Perrin et al., 2001, presents a similar finding). This shows a critical weakness of model comparison studies that use a small number of basins: Results are conditional on the choice of catchments and thus very difficult to generalize to other places.

Large-sample studies allow patterns to emerge, of which Figure 9a provides an example: Some model structures seem relatively unsuitable for flow simulation with the KGE(Q) objective function, but of the models that do tend to rank better on this objective, models with more parameters appear to have more flexibility and tend to rank better in larger numbers of catchments than models with fewer parameters. This leaves modelers working with conceptual models in large numbers of catchments facing a dilemma: Parsimonious models are preferable because their parameters will be better identifiable (e.g., Jakeman & Hornberger, 1993; Nash & Sutcliffe, 1970; Wagener et al., 2003), but, as our results suggest, models with more parameters seem to have the flexibility to accurately reproduce hydrographs in a much wider range of catchments without any obvious risk of being overfitted to the calibration data.

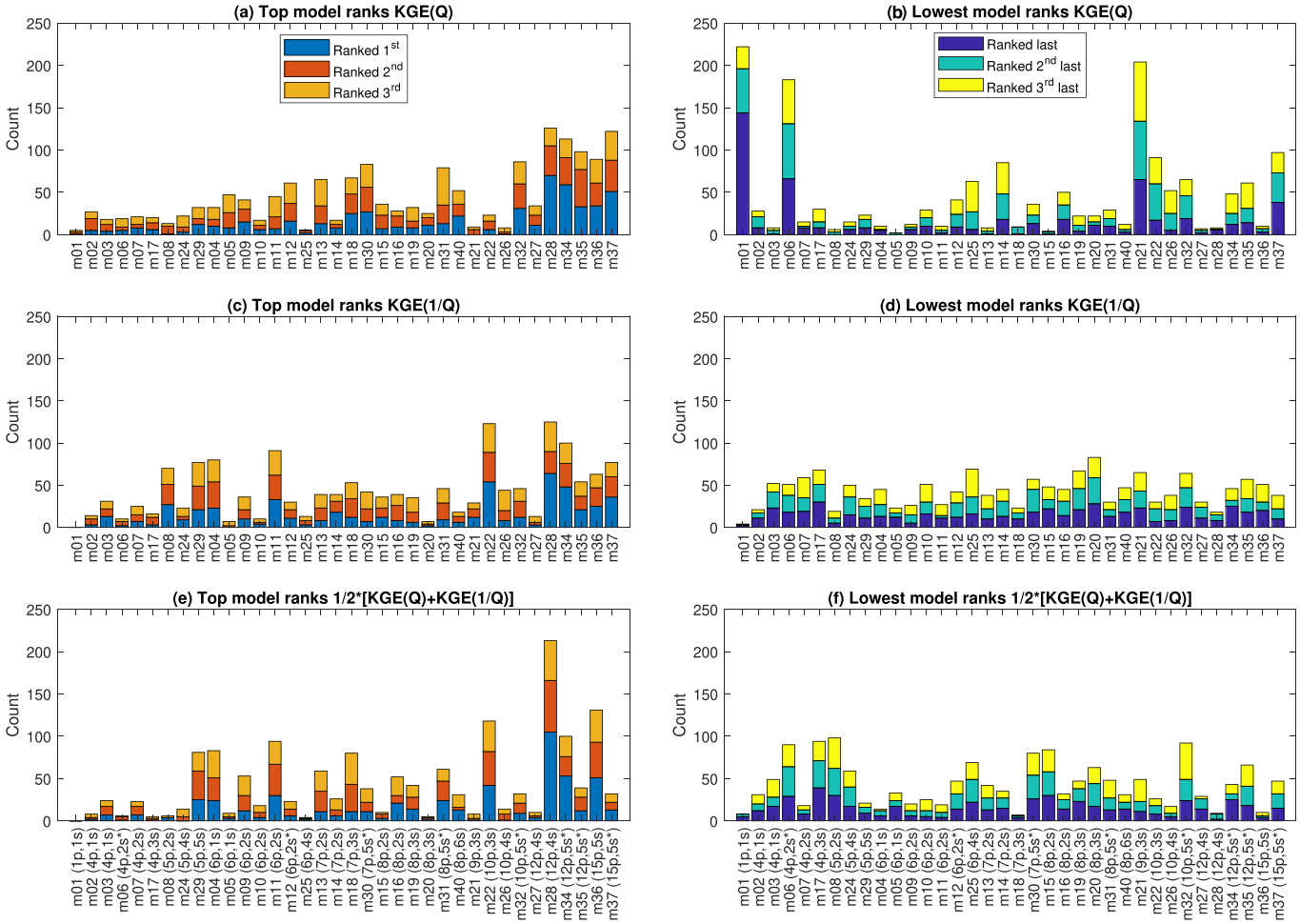

Figure 9. Overview of how often each model ranks (a, c, e) in the top three and (b, d, f) in the bottom three models during evaluation. Note that these instances are only counted in cases where the model at least beats the catchment-specific benchmark. For example, there are only 548 counts of any model being ranked first on the KGE(Q) objective, while our sample contains 559 catchments, because in 11 catchments not a single model beats the benchmark and in these catchments no model is assigned a rank. Models that include a snow module are indicated with an asterisk (*).

Last, we attempt to answer "How does relative model performance relate to known catchment attributes?" We found the strongest relation between relative model rank and observed streamflow signatures, while relations with climatic attributes and catchment descriptors were less clear (Figures 7, S21, and S22). The lack of correlation with some attributes may be due to low information content in these attributes but might also be the results of our experimental design. Given that models were calibrated to streamflow data, it is perhaps not surprising to see model performance correlate most strongly with streamflow signatures, because the models are naturally sorted by their ability to replicate certain regimes as a consequence of the calibration procedure. It is possible that the relation between model performance and other catchment attributes becomes clearer when those attributes are specifically used during calibration (e.g., calibrating against streamflow, evaporation and soil moisture observations simultaneously might clarify relations between model performance and climate and/or soil characteristics), but such calibration approaches also carry new challenges with them such as the commensurability between model states and real-world observations. Currently, our findings reinforce the idea that certain model structures are better suited for simulation of certain flow regimes and our results suggest that models that share certain structural elements show similar suitability for certain flow regimes. Our results give rise to several hypotheses about conceptual model behavior (see section 4.4.2 and the note on Model m28 in this section), but all were formulated after we calibrated, evaluated, ranked, and grouped the models. Strict testing of the hypotheses is thus necessary (see, e.g., Beven, 2000, 2018; Clark et al., 2011; Fenicia et al., 2014; Kirchner, 2006; Pfister & Kirchner, 2017) before these ideas can be used to guide model development.

5.2. The Need to Select an Appropriate Hydrological Model

This study provides large-sample evidence for the need for more thorough and process-based model evaluations (e.g., Clark et al., 2016; Gupta et al., 2012; Kirchner, 2006; Wrede et al., 2015). There are too many models that are superficially similar in terms of efficiency scores but internally different in terms of process representation. To increase hydrologic understanding and generate robust long-term projections of future water resources, more effort needs to be devoted to understanding (1) which hydrologic processes are dominant where, (2) which model structures contain appropriate representations of these dominant processes and should thus be used in a given catchment (for a given definition of "appropriate"), and (3) how dominant hydrologic processes and consequently the criteria for what constitutes an "appropriate" model might change in the future in a given catchment. Only with such understanding can we confidently select a model structure for a given catchment and study purpose.

It has been long known that models with fewer calibrated parameters can compete with more parameter -heavy models in terms of model performance (e.g., Jakeman & Hornberger, 1993; Perrin et al., 2001). Overfitting and the inability to properly identify parameter values through calibration to streamflow data are often cited as a reason for this (e.g., Beven, 1989; Kuczera & Mroczkowski, 1998). Parsimonious models with few calibration parameters are often preferred over more complex models to avoid these issues. However, such simple models cannot contain all potentially relevant hydrologic processes, because this would require more parameters than can be identified from streamflow data alone. This leads to a dilemma succinctly stated in Kuczera and Mroczkowski (1998): "A simple model cannot be relied upon to make meaningful extrapolative predictions, whereas a complex model may have the potential but because of information constraints may be unable to realize it." Yet, such complex models are required if predictions under changing conditions are to be made (Kirchner, 2006). Our results suggest that models with a larger number of parameters (up to 15 in this study) are less vulnerable to parameter overfitting than might be expected (although parameter identifiability might remain an issue; see section 5.4). Therefore, when a choice must be made between several model structures for prediction under changing conditions there is no clear justification for selecting the model with the fewest calibration parameters as the preferable alternative (cf. Oreskes et al., 1994; Reichert & Omlin, 1997). Instead, analysis of the dominant hydrologic processes, possible changes in these processes and each models' ability to reflect both current and future processes should form the core of such a decision.

5.3. How Representative Are the Catchment and Model Sample?

We have remarked on the fact that results from modeling studies are conditional in the sample of models and catchments used. We therefore briefly summarize our findings about the representativeness of our catchment and model sample and refer the interested reader to supporting information Text S3 for more details.

We use the hydrological climate classification of Knoben, Woods, and Freer (2018) to quantify where the CAMELS catchments fall in relation to the global distribution of hydroclimates. This classification uses three axes that describe annual average aridity, the within-year variability in the water-energy balance and the fraction of precipitation that occurs as snow. There are few CAMELS catchments on either end of the aridity scale, few catchments with low within-year aridity seasonality, and snow-dominated CAMELS catchments cover a fairly narrow range out of all possible snow-dominated conditions. In geographical terms, care should be taken when extrapolating our findings to climates with more extreme aridity values (e.g., deserts and tropical rain forests), to regions with less seasonally varied aridity values (e.g., climatic transition zones on the edges of deserts and rain forest), and to places with a low mean temperatures combined with a less pronounced summer-winter temperature cycle (e.g., taiga).

It is commonly assumed that models in an ensemble are sufficiently varied if the model simulations bracket the observations (Clark et al., 2008). This is the case for the majority of catchments during evaluation. Clear exceptions are mountainous snow-dominated catchments and several catchments on the Pacific Northwest, which can indicate a lack of diversity in our model ensemble (likely the case for snow-dominated catchments) or the presence of bias in the forcing data (a likely explanation for the Pacific Northwest). Despite mostly bracketing the observations, the model ensemble shows a strong seasonal bias with a tendency toward underprediction in late spring and summer and overprediction in late autumn and winter. This could be due to the spacing of the MARRMoT models in the overall model space. Skewed sampling of

model structures can bias model comparison studies but the extent to which this is an issue is currently difficult to quantify. Metrics that define model similarity and model spread in the total model space are needed to address this question in more depth.

5.4. Study Limitations

This section briefly describes various limitations in the current study set up and possible ways to address these in future work. First, our analyses are mostly based on general performance metrics that aggregate model performance into a single efficiency score. Higher resolution diagnostics such as seasonal or time step based performance metrics (see, e.g., Coxon et al., 2014), or signature-based calibration and evaluation (see, e.g., Westerberg et al., 2014) might provide insight into why there is considerable equifinality in aggregated model performance across our sample.

Second, we use lumped models, catchment-averaged daily forcing data and average catchment attributes. Spatial heterogeneity is not accounted for beyond parametrizations of contributing catchment area in certain model structures, in an attempt to keep the analysis manageable. This lack of spatial explicitness makes it challenging to get clear answers to questions that require more nuanced analysis of the hydrograph and/or detailed process consideration, and much work remains to be done.

Third, this work focuses on model structure uncertainty and leaves data and parameter uncertainty mostly unaccounted for. Our experimental design is constrained by a need to limit computational times but ignoring data uncertainty (see, e.g., McMillan et al., 2012, 2018) can force the calibration procedure to compensate for errors in the measured rainfall-runoff relationship and thus influence results. Equally, although we see no evidence of parameter overfitting, it is still possible that parameters are poorly identifiable. We have performed a short investigation of the impact of using only a single parameter set per model per catchment and believe that patterns across all catchments and models can be inferred from this approach (see supporting information Text S4 for details). We caution against using our data to investigate a single catchment without accounting for parameter uncertainty, because differences between calibrated parameter sets can be substantial in a given basin, even if patterns across the sample of all basins remain relatively stable. Our results do suggest that models can be divided into groups, where models within each group are similarly suited toward particular flow regimes. These groups can be used to select a small number of promising models within our ensemble, with each selected model being representative of several others. This reduces the computational load and is a potential way to allow future studies more room to account for data and parameter uncertainties.

5.5. Fostering Further Work

The computational demands of studies such as this can be high. To facilitate further research, calibration results of all models (parameters, simulated model storages and fluxes and obtained efficiency values) are made available on the University of Bristol data repository (dx.doi.org/10.5523/bris.2zutxh2qeep6y2cy6scwgk9eqj). The CAMELS data set and MARRMoT modeling toolbox are also freely available and can be found through their respective references.

6. Conclusions

We calibrated 36 lumped conceptual models for streamflow simulation in 559 catchments across the United States, using three different formulations of the KGE as objective functions. We used a benchmark based on the mean or median calendar day flow (depending on the catchment) to define a baseline of expected model performance for each catchment. This benchmark proved hardest to beat in mountainous snow-dominated catchments: In 11 of these catchments no model managed to beat the benchmark, indicating that persistent features of the hydrograph are systematically poorly simulated in these places. In wet nonsnowy catchments, the majority of models managed to beat the benchmark. In arid catchments model choice seemed to matter most: Models do exist that beat the benchmark (and by similar margins as models in wetter catchments do), but these must be carefully selected. In nearly all catchments model equifinality can be high. For approximately 500 catchments, between 1 and up to 28 models can be within 0.05 KGE from the best model in each catchment. Our results indicate that there is little relation between model performance and number of parameters and there is no evidence of increased risk of overfitting of models with more parameters compared to models with fewer parameters. Instead, our results suggest that the choice of model

parametrization (i.e., which equations are used and how parameters are used within them) is more important to dictate its suitability for flow simulation with a given objective function and the flow regimes the model is capable of simulating well. In fact, our results suggest that if the model is suitable for a given objective function, models with more parameters tend to have increased flexibility compared to models with fewer parameters. This flexibility allows them to perform well outside the calibration period in larger numbers of catchments. It remains difficult to explain the type of catchments where a model might do well with attributes that quantify the catchments' geologic, topographic, soil, and vegetation attributes. Instead, model suitability seems to relate strongest to the streamflow regime each catchment generates, and we show an initial assessment that relates commonalities in model structure to similarities in model performance. Given our catchment-averaged approach to model use, data, and analysis and the fact that our hypotheses about model structure similarity were formulated after we calibrated and evaluated our models, more detailed investigation of between-model differences is needed, and care should be taken when applying our findings to future modeling efforts that extend beyond the limits of our approach.

Data Availability Statement

CAMELS data can be downloaded from https://ncar.github.io/hydrology/datasets/CAMELS_attributes (Addor et al., 2017). The latest MARRMoT model code and supporting information can be downloaded from https://github.com/wknoben/MARRMoT (Knoben et al., 2018c); MARRMoT v1.0 which was used for this work is available online (from https://dx.doi.org/10.5281/zenodo.2482542). A data package containing calibrated parameter values for the models used in this work, obtained efficiency values, and time series of simulated flows, internal fluxes, and model states can be downloaded from the University of Bristol data repository (dx.doi.org/10.5523/bris.2zutxh2qeep6y2cy6scwgk9eqj).

Acknowledgments
We are grateful for the detailed comments provided by the Editors, James Craig, and three anonymous reviewers, which have helped us substantially improve this paper. This work was funded by the EPSRC WISE CDT, Grant Reference EP/L016214/1. W. K.'s visit to the University of Melbourne was cofunded by the Melbourne School of Engineering Visiting Fellows scheme. M. P. is the recipient of an Australian Research Council Future Fellowship (FT120100130).

References

Addor, N., Do, H. X., Alvarez-Garreton, C., Coxon, G., Fowler, K., & Mendoza, P. A. (2019). Large-sample hydrology: Recent progress, guidelines for new datasets and grand challenges. *Hydrological Sciences Journal*, *65*, 712–725. https://doi.org/10.1080/02626667.2019.1683182

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

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*, 5293–5313. https://doi.org/10.5194/hess-2017-169

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., et al. (2018). The CAMELS-CL dataset: Catchment attributes and meteorology for large sample studies—Chile dataset. *Hydrology and Earth System Sciences*, *22*(11), 5817–5846. https://doi.org/10.5194/hess-22-5817-2018

Ambroise, B., Freer, J., & Beven, K. (1996). Application of a generalized TOPMODEL to the small Ringelbach catchment, Vosges, France. *Water Resources Research*, *32*(7), 2147–2159. https://doi.org/10.1029/95WR03715

Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., et al. (2009). Crash tests for a standardized evaluation of hydrological models. *Hydrology and Earth System Sciences*, *13*(10), 1757–1764. https://doi.org/10.5194/hess-13-1757-2009

Arsenault, R., Poulin, A., Côté, P., & Brissette, F. (2014). Comparison of stochastic optimization algorithms in hydrological model calibration. *Journal of Hydrologic Engineering*, *19*(7), 1374–1384. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000938

Bell, V. A., Carrington, D. S., & Moore, R. J. (2001). Comparison of rainfall-runoff models for flood forecasting—Part 2: Environment agency.

Berghuijs, W. R., Sivapalan, M., Woods, R. A., & Savenije, H. H. G. (2014). Patterns of similarity of seasonal water balances: A window into streamflow variability over a range of time scales. *Water Resources Research*, *50*, 5638–5661. https://doi.org/10.1002/2014WR015692

Beven, K. (1989). Changing ideas in hydrology—The case of physically-based models. *Journal of Hydrology*, *105*(1-2), 157–172. https://doi.org/10.1016/0022-1694(89)90101-7

Beven, K. (2000). Uniqueness of place and process representations in hydrological modelling. https://doi.org/10.5194/hess-4-203-2000

Beven, K. (2012). *Rainfall-runoff modelling: The primer* (2nd ed.).: John Wiley and Sons Ltd.

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

Biondi, D., Freni, G., Iacobellis, V., Mascaro, G., & Montanari, A. (2012). Validation of hydrological models: Conceptual basis, methodological approaches and a proposal for a code of practice. *Physics and Chemistry of the Earth, Parts A/B/C*, *42-44*, 70–76. https://doi.org/10.1016/j.pce.2011.07.037

Budyko, M. I. (1974). *Climate and life*: New York: Academic Press.

Ceola, S., Arheimer, B., Baratti, E., Blöschl, G., Capell, R., Castellarin, A., et al. (2015). Virtual laboratories: New opportunities for collaborative water science. *Hydrology and Earth System Sciences*, *19*, 2101–2117. https://doi.org/10.5194/hess-19-2101-2015

Chiew, F. H. S. (1990). Estimating groundwater recharge using an integrated surface and groundwater model (PhD thesis), University of Melbourne.

Chiew, F., & McMahon, T. (1994). Application of the daily rainfall-runoff model MODHYDROLOG to 28 Australian catchments. *Journal of Hydrology*, *153*(1-4), 383–416. https://doi.org/10.1016/0022-1694(94)90200-3

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., Lundquist, J. D., Kavetski, D., Rupp, D. E., Woods, R. A., et al. (2015). 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., Schaefli, B., Schymanski, S. J., Samaniego, L., Luce, C. H., Jackson, B. M., et al. (2016). Improving the theoretical underpinnings of process-based hydrologic models. *Water Resources Research*, *52*, 2350–2365. https://doi.org/10.1002/2015WR017910

Clark, M. P., Slater, A. G., Rupp, 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

Coron, L., Andréassian, V., Perrin, C., Lerat, J., Vaze, J., Bourqui, M., & Hendrickx, F. (2012). Crash testing hydrological models in contrasted climate conditions: An experiment on 216 Australian catchments. *Water Resources Research*, *48*, W05552. https://doi.org/10.1029/2011WR011721

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

Dakhlaoui, H., Ruelland, D., Tramblay, Y., & Bargaoui, Z. (2017). Evaluating the robustness of conceptual rainfall-runoff models under climate variability in northern Tunisia. *Journal of Hydrology*, *550*, 201–217. https://doi.org/10.1016/j.jhydrol.2017.04.032

de Boer-Euser, T., Bouaziz, L., De Niel, J., Brauer, C., Dewals, B., Drogue, 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

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

Fenicia, F., Kavetski, D., & Savenije, H. H. G. (2011). Elements of a flexible approach for conceptual hydrological modeling: 1. Motivation and theoretical development. *Water Resources Research*, *47*, W11510. https://doi.org/10.1029/2010WR010174

Fenicia, F., Kavetski, D., Savenije, H. H. G., Clark, M. P., Schoups, G., Pfister, L., & 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.9726

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

Fenicia, F., Savenije, H. H. G., Matgen, P., & Pfister, L. (2008). Understanding catchment behavior through stepwise model concept improvement. *Water Resources Research*, *44*, W01402. https://doi.org/10.1029/2006WR005563

Fowler, K., Peel, M., Western, A., & Zhang, L. (2018). Improved rainfall-runoff calibration for drying climate: Choice of objective function. *Water Resources Research*, *54*, 3392–3408. https://doi.org/10.1029/2017WR022466

Franchini, M., & Pacciani, M. (1991). Comparative analysis of several conceptual rainfall-runoff models. *Journal of Hydrology*, *122*(1-4), 161–219. https://doi.org/10.1016/0022-1694(91)90178-K

Garrick, M., Cunnane, C., & Nash, J. E. (1978). A criterion of efficiency for rainfall-runoff models. *Journal of Hydrology*, *36*(3-4), 375–381. https://doi.org/10.1016/0022-1694(78)90155-5

Grayson, R. B., & Blöschl, G. (Eds.) (2001). *Spatial patterns in catchment hydrology: Observations and modelling Edited by* Grayson, R. B., & Blöschl, G., pp. 416: Cambridge University Press.

Gupta, H. V., Clark, M. P., Vrugt, J., Abramowitz, G., & Ye, M. (2012). Towards a comprehensive assessment of model structural adequacy. *Water Resources Research*, *48*, W08301. https://doi.org/10.1029/2011WR011044

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Höge, M., Wöhling, T., & Nowak, W. (2018). A primer for model selection: The decisive role of model complexity. *Water Resources Research*, *54*, 1688–1715. https://doi.org/10.1002/2017WR021902

Hansen, N. (2009). References to CMA-ES applications. https://www.cmap.polytechnique.fr/∼nikolaus.hansen/cmaapplications.pdf

Hansen, N. (2016). The CMA evolution strategy: A tutorial.

Hansen, N., Auger, A., Ros, R., Finck, S., & Pošík, P. (2010). Comparing results of 31 algorithms from the Black-Box Optimization Benchmarking BBOB-2009. In *Workshop Proceedings of the GECCO Genetic and Evolutionary Computation Conference 2010*, pp. 1689–1696.

Hansen, N., Müller, S. D., & Koumoutsakos, P. (2003). Reducing the time complexity of the derandomized Evolution Strategy with Covariance Matrix Adaptation (CMA-ES). *Evolutionary Computation*, *11*(1), 1–18. https://doi.org/10.1162/106365603321828970

Hansen, N., & Ostermeier, A. (1996). Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In *Proceedings of the IEEE Conference on Evolutionary Computation*, pp. 312–317. https://doi.org/10.1109/icec.1996.542381

Hansen, N., & Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, *9*(2), 159–195. https://doi.org/10.1162/106365601750190398

Hogue, T. S., Bastidas, L. A., Gupta, H. V., & Sorooshian, S. (2006). Evaluating model performance and parameter behavior for varying levels of land surface model complexity. *Water Resources Research*, *42*, W08430. https://doi.org/10.1029/2005WR004440

Hrachowitz, M., & Clark, M. P. (2017). HESS Opinions: The complementary merits of competing modelling philosophies in hydrology. *Hydrology and Earth System Sciences*, *21*, 3953–3973. https://doi.org/10.5194/hess-21-3953-2017

Jakeman, A. J., & Hornberger, G. M. (1993). How much complexity is warranted in a rainfall-runoff model? *Water Resources Research*, *29*(8), 2637–2649. https://doi.org/10.1029/93WR00877

Jayawardena, A. W., & Zhou, M. C. (2000). A modified spatial soil moisture storage capacity distribution curve for the Xinanjiang model. *Journal of Hydrology*, *227*, 93–113. https://doi.org/10.1016/S0022-1694(99)00173-0

Kavetski, D., Kuczera, G., & Franks, S. W. (2006). Calibration of conceptual hydrological models revisited: 1. Overcoming numerical artefacts. *Journal of Hydrology*, *320*(1–2), 173–186. https://doi.org/10.1016/j.jhydrol.2005.07.012

Kim, K. B., Kwon, H. H., & Han, D. (2018). Exploration of warm-up period in conceptual hydrological modelling. *Journal of Hydrology*, *556*, 194–210. https://doi.org/10.1016/j.jhydrol.2017.11.015

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

Knoben, W. J. M., Freer, J. E., Fowler, K. J. A., Peel, M. C., & Woods, R. A. (2018a). Modular Assessment of Rainfall-Runoff Models Toolbox v1.0—Matlab code for 46 conceptual hydrologic models. https://doi.org/10.5281/zenodo.2482542

Knoben, W. J. M., Freer, J. E., Fowler, K. J. A., Peel, M. C., & Woods, R. A. (2019). Modular Assessment of Rainfall-Runoff Models Toolbox (MARRMoT) v1.2: An open-source, extendable framework providing implementations of 46 conceptual hydrologic models as continuous state-space formulations. *Geoscientific Model Development*, *12*, 2463–2480. https://doi.org/10.5194/gmd-2018-332

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Technical note: Inherent benchmark or not? Comparing Nash Sutcliffe and Kling Gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331. https://doi.org/10.5194/hess-23-4323-2019

Knoben, W. J. M., Woods, R. A., & Freer, J. E. (2018). A quantitative hydrological climate classification evaluated with independent streamflow data. *Water Resources Research*, *54*, 5088–5109. https://doi.org/10.1029/2018WR022913

Knoben, W. J. M., Woods, R. A., & Freer, J. E. (2018c). Climate data from paper "A quantitative hydrological climate classification evaluated with independent stream flow data". https://doi.org/10.5523/bris.16ctquxqxk46h2v61gz7drcdz3

Kraft, P., Vaché, K. B., Frede, H.-G., & Breuer, L. (2011). CMF: 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

Krueger, T., Freer, J., Quinton, J. N., Macleod, 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

Krueger, T., Freer, J., Quinton, J. N., Macleod, 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

Krysanova, V., Vetter, T., Eisner, S., Huang, S., Pechlivanidis, I., Strauch, M., et al. (2017). Intercomparison of regional-scale hydrological models and climate change impacts projected for 12 large river basins worldwide—A synthesis. *Environmental Research Letters*, *12*(10), 105,002. https://doi.org/10.1088/1748-9326/aa8359

Kuczera, G., & Mroczkowski, M. (1998). Assessment of hydrologic parameter uncertainty and the worth of multiresponse data. *Water Resources Research*, *34*(6), 1481–1489. https://doi.org/10.1029/98WR00496

Kuentz, A., Arheimer, B., Hundecha, Y., & Wagener, T. (2017). Understanding hydrologic variability across Europe through catchment classification. *Hydrology and Earth System Sciences*, *21*(6), 2863–2879. https://doi.org/10.5194/hess-21-2863-2017

Lane, R. A., Coxon, G., Freer, J. E., Wagener, T., Johnes, P. J., Bloomfield, J. P., et al. (2019). Benchmarking the predictive capability of hydrological models for river flow and flood peak predictions across over 1000 catchments in Great Britain. *Hydrology and Earth System Sciences*, *23*(10), 4011–4032. https://doi.org/10.5194/hess-23-4011-2019

Lidén, R., & Harlin, J. (2000). Analysis of conceptual rainfall-runoff modelling performance in different climates. *Journal of Hydrology*, *238*(3-4), 231–247. https://doi.org/10.1016/S0022-1694(00)00330-9

Linsley, R. K. (1982). Rainfall-runoff models—An overview. In Singh, V. P. (Ed.), *Rainfall-runoff relationship* (pp. 582): Water Resources Publications, USA.

Lute, A. C., & Luce, C. H. (2017). Are model transferability and complexity antithetical? Insights from validation of a variable-complexity empirical snow model in space and time. *Water Resources Research*, *53*, 8825–8850. https://doi.org/10.1002/2017WR020752

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, *18*(1), 50–60. https://doi.org/10.1214/aoms/1177730491

Martinez, G. F., & Gupta, H. V. (2011). Hydrologic consistency as a basis for assessing complexity of monthly water balance models for the continental United States. *Water Resources Research*, *47*, W12540. https://doi.org/10.1029/2011WR011229

McGlynn, B. L., McDonnel, J. J., & Brammer, D. D. (2002). A review of the evolving perceptual model of hillslope flowpaths at the Maimai catchments, New Zealand. *Journal of Hydrology*, *257*, 1–26.

McMillan, H., Krueger, T., & Freer, J. (2012). Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality. *Hydrological Processes*, *26*(26), 4078–4111. https://doi.org/10.1002/hyp.9384

McMillan, H. K., Westerberg, I. K., & Krueger, T. (2018). Hydrological data uncertainty and its implications. *Wiley Interdisciplinary Reviews: Water*, *5*(6), e1319. https://doi.org/10.1002/wat2.1319

Melsen, L. A., Addor, N., Mizukami, N., Newman, A. J., Torfs, P. J. J. F., Clark, M. P., et al. (2018). Mapping (dis)agreement in hydrologic projections. *Hydrology and Earth System Sciences*, *22*(3), 1775–1791. https://doi.org/10.5194/hess-22-1775-2018

Merz, R., Parajka, J., & Blöschl, G. (2011). Time stability of catchment model parameters: Implications for climate impact analyses. *Water Resources Research*, *47*, W02531. https://doi.org/10.1029/2010WR009505

Moore, R. J., & Bell, V. A. (2001). Comparison of rainfall-runoff models for flood forecasting. Part 1: Literature review of models. Bristol: Environment Agency.

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models Part I—A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., et al. (2015). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209–223. https://doi.org/10.5194/hess-19-209-2015

Newman, A. J., Mizukami, N., Clark, M. P., Wood, A. W., Nijssen, B., & Nearing, G. (2017). Benchmarking of a physically based hydrologic model. *Journal of Hydrometeorology*, *18*(8), 2215–2225. https://doi.org/10.1175/JHM-D-16-0284.1

Nijzink, R., Hutton, C., Pechlivanidis, I., Capell, R., Arheimer, B., Freer, J., et al. (2016). The evolution of root-zone moisture capacities after deforestation: A step towards hydrological predictions under change?. *Hydrology and Earth System Sciences*, *20*(12), 4775–4799. https://doi.org/10.5194/hess-20-4775-2016

Oreskes, N., Shrader-Frechette, K., & Belitz, K. (1994). Verification, validation, and confirmation of numerical models in the Earth sciences. *Science*, *263*(5147), 641–646. https://doi.org/10.1126/science.263.5147.641

Oudin, L., Salavati, B., Furusho-percot, C., Ribstein, P., & Saadi, M. (2018). Hydrological impacts of urbanization at the catchment scale. *Journal of Hydrology*, *559*, 774–786. https://doi.org/10.1016/j.jhydrol.2018.02.064

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

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

Peters, N. E., Freer, J., & Beven, K. (2003). Modelling hydrologic responses in a small forested catchment (Panola Mountain, Georgia, USA): A comparison of the original and a new dynamic TOPMODEL. *Hydrological Processes*, *17*(2), 345–362. https://doi.org/10.1002/hyp.1128

Peterson, T. J., & Western, A. W. (2014). Nonlinear time-series modeling of unconfined groundwater head. *Water Resources Research*, *50*, 8330–8355. https://doi.org/10.1002/2013WR014800

Pfister, L., & Kirchner, J. W. (2017). Debates hypothesis testing in hydrology: Theory and practice. *Water Resources Research*, *53*, 1792–1798. https://doi.org/10.1002/2016WR020116
Pilgrim, D. H., Chapman, T. G., & Doran, D. G. (1988). Problems of rainfall-runoff modelling in arid and semiarid regions. *Hydrological Sciences Journal*, *33*(4), 379–400. https://doi.org/10.1080/02626668809491261
Priestley, C. H. B., & Taylor, R. J. (1972). On the assessment of surface heat flux and evaporation using large-scale parameters. *Monthly Weather Review*, *100*(2), 81–92. https://doi.org/10.1175/1520-0493(1972)100<0081:OTAOSH>2.3.CO;2
Pushpalatha, R., Perrin, C., Moine, N. L., & Andréassian, V. (2012). A review of efficiency criteria suitable for evaluating low-flow simulations. *Journal of Hydrology*, *420-421*, 171–182. https://doi.org/10.1016/j.jhydrol.2011.11.055
Reichert, P., & Omlin, M. (1997). On the usefulness of overparameterized ecological models. *Ecological Modelling*, *95*(2-3), 289–299. https://doi.org/10.1016/S0304-3800(96)00043-9
Santos, L., Thirel, G., & Perrin, C. (2018). Technical note: Pitfalls in using log-transformed flows within the KGE criterion. *Hydrology and Earth System Sciences*, *22*(8), 4583–4591. https://doi.org/10.5194/hess-22-4583-2018
Savenije, H. H. G. (2009). HESS opinions "The art of hydrology". *Hydrology and Earth System Sciences*, *13*(2), 157–161. https://doi.org/10.5194/hess-13-157-2009
Schaefli, B., & Gupta, H. V. (2007). Do Nash values have value?. *Hydrological Processes*, *21*, 2075–2080. https://doi.org/10.1002/hyp.6825
Schoups, G., Van De Giesen, N. C., & Savenije, H. H. G. (2008). Model complexity control for hydrologic prediction. *Water Resources Research*, *44*, W00B03. https://doi.org/10.1029/2008WR006836
Schoups, G., Vrugt, J. A., Fenicia, F., & Van De Giesen, N. C. (2010). Corruption of accuracy and efficiency of Markov chain Monte Carlo simulation by inaccurate numerical implementation of conceptual hydrologic models. *Water Resources Research*, *46*, W10530. https://doi.org/10.1029/2009WR008648
Seibert, J. (2001). On the need for benchmarks in hydrological modelling. *Hydrological Processes*, *15*(6), 1063–1064.
Seibert, J., Vis, M. J. P., Lewis, E., & van Meerveld, H. J. (2018). Upper and lower benchmarks in hydrological modelling. *Hydrological Processes*, *32*(8), 1120–1125. https://doi.org/10.1002/hyp.11476
Seiller, G., Anctil, F., & Perrin, C. (2012). Multimodel evaluation of twenty lumped hydrological models under contrasted climate conditions. *Hydrology and Earth System Sciences*, *16*(4), 1171–1189. https://doi.org/10.5194/hess-16-1171-2012
Shaw, E. M., Beven, K. J., Chappell, N. A., & Lamb, R. (2011). *Hydrology in practice* (4th ed.).: Spon Press.
Singh, V. P., & Woolhiser, D. A. (2002). Mathematical modeling of watershed hydrology. *Journal of Hydrologic Engineering*, *7*(4), 270–292. https://doi.org/10.1061/(ASCE)1084-0699(2002)7:4(270)
Sittner, W. T. (1976). WMO project on intercomparison of conceptual models used in hydrological forecasting. *Hydrological Sciences Bulletin*, *21*(1), 203–213. https://doi.org/10.1080/02626667609491617
US Geological Survey (2018). National Water Information System: Web interface. https://nwis.waterdata.usgs.gov/nwis/dv, .
Van Esse, W. R., Perrin, C., Booij, M. J., Augustijn, D. C. M., Fenicia, 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
Wagener, T., McIntyre, N., Lees, M. J. J., Wheater, H. S. S., & Gupta, H. V. V. (2003). Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis. *Hydrological Processes*, *17*(2), 455–476. https://doi.org/10.1002/hyp.1135
Westerberg, I. K., Gong, L., Beven, K. J., Seibert, J., Semedo, A., Xu, C. Y., & Halldin, S. (2014). Regional water balance modelling using flow-duration curves with observational uncertainties. *Hydrology and Earth System Sciences*, *18*(8), 2993–3013. https://doi.org/10.5194/hess-18-2993-2014
Winter, C. L., & Nychka, D. (2010). Forecasting skill of model averages. *Stochastic Environmental Research and Risk Assessment*, *24*(5), 633–638. https://doi.org/10.1007/s00477-009-0350-y
Wrede, S., Fenicia, F., Martínez-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
Zhao, R.-J. (1992). The Xinanjiang model applied in China. *Journal of Hydrology*, *135*(1-4), 371–381. https://doi.org/10.1016/0022-1694(92)90096-E

References From the Supporting Information

Allen, R., Pereira, L., Raes, D., & Smith, M. (1998). *Crop evapotranspiration—Guidelines for computing crop water requirements—FAO Irrigation and drainage paper 56*, pp. 15. Rome.
McMahon, T. A., Peel, M. C., Lowe, L., Srikanthan, R., & McVicar, T. R. (2013). Estimating actual, potential, reference crop and pan evaporation using standard meteorological data: A pragmatic synthesis—Supplement. *Hydrology and Earth System Sciences*, *17*(4), 1331–1363. https://doi.org/10.5194/hess-17-1331-2013