Teaching hydrological modelling: illustrating model structure uncertainty with a ready-to-use computational exercise

Wouter J. M. Knoben¹ and Diana Spieler²
¹Coldwater Laboratory, University of Saskatchewan, Canmore, Alberta, Canada
²Institute of Hydrology and Meteorology, Technische Universität Dresden, Dresden, Germany

Correspondence: Diana Spieler (diana.spieler@tu-dresden.de)

Received: 16 January 2021 – Discussion started: 21 January 2021
Revised: 31 March 2022 – Accepted: 16 May 2022 – Published: 29 June 2022

Abstract. Estimating the impact of different sources of uncertainty along the modelling chain is an important skill graduates are expected to have. Broadly speaking, educators can cover uncertainty in hydrological modelling by differentiating uncertainty in data, model parameters and model structure. This provides students with insights on the impact of uncertainties on modelling results and thus on the usability of the acquired model simulations for decision making. A survey among teachers in the Earth and environmental sciences showed that model structural uncertainty is the least represented uncertainty group in teaching. This paper introduces a computational exercise that introduces students to the basics of model structure uncertainty through two ready-to-use modelling experiments. These experiments require either Matlab or Octave, and use the open-source Modular Assessment of Rainfall-Runoff Models Toolbox (MARRMoT) and the open-source Catchment Attributes and Meteorology for Large-sample Studies (CAMELS) data set. The exercise is short and can easily be integrated into an existing hydrological curriculum, with only a limited time investment needed to introduce the topic of model structure uncertainty and run the exercise. Two trial applications at the Technische Universität Dresden (Germany) showed that the exercise can be completed in two afternoons or four 90 min sessions and that the provided setup effectively transfers the intended insights about model structure uncertainty.

1 Introduction

The ability to use computer models to provide hydrological predictions is a critical skill for young hydrologists (Seibert et al., 2013; Wagener and McIntyre, 2007). Model use is so widespread that students will have to generate, use or present modelling results at some point in their professional career (Seibert et al., 2013). A very wide range of different models currently exists and it is arguably less important for students to learn how to use any specific model than to be taught general modelling concepts. Students should have some understanding of different modelling philosophies, learn to use different model types and be aware of the strengths and limitations of hydrological modelling (Wagener et al., 2012). Given the societal need to provide hydrological predictions far into the future and the unknown (Kirchner, 2006), a core competence for young professionals is knowing how to provide such predictions in a scientifically sound manner.

Understanding of uncertainty in the modelling process is key to interpreting model results (e.g. Pechlivanidis et al., 2011; Blöschl and Montanari, 2010; Beven et al., 2011; Mendoza et al., 2015, among many others). Modelling uncertainties can be roughly classified as relating to the input and evaluation data, the estimation or calibration of model parameters and the choice of equations that make up the model structure. These concepts should be an integral part of the hydrological curriculum (Wagener et al., 2012; AghaKouchak et al., 2013; Thompson et al., 2012) in a teaching structure that includes student-driven, hands-on exercises that reinforce the taught concepts (Thompson et al., 2012). A survey among 101 teachers in the Earth and environmental sciences (see Sect. S1 in the Supplement) shows large differences in how much time is spent on teaching hydrological modelling in general, whether model-related uncertainty is part of the course and, if so, which aspects of uncertainty are covered. Based on the survey, model structural uncertainty is the least
represented uncertainty aspect in teaching. The main reason
named for not covering model-related uncertainty is a lack of
time, while the lack of good teaching materials is the second-
most common explanation. Just 6% of respondents that did
not cover uncertainty in their classes stated that the topic
would be covered in another course.

Selecting a model that faithfully represents current and
future hydrological conditions in a given catchment is crit-
ical for realistic long-term projections of water availability.
In other words, one requires “the right answers for the right
reasons” (Kirchner, 2006). The difficult task of finding an
appropriate model structure, i.e., the combination of which
hydrological processes are included in a model, which equa-
tions are used to describe these processes and how model
states and fluxes are connected, can be referred to as model
structure uncertainty and is a significant source of overall
modelling uncertainty (Di Baldassarre and Montanari, 2009).

Model structure uncertainty is being investigated with in-
creasing numbers of models in increasingly varied selections
of catchments (e.g. Perrin et al., 2001; Butts et al., 2004;
Duan et al., 2006; van Esse et al., 2013; Knoben et al., 2020;
Spieler et al., 2020) and results are consistent: model choice
matters and selecting an inappropriate model for a given
catchment can lead to simulations of questionable quality.
Regrettably, suitability of a given model for the task at hand
is not always the main driver in model selection. Prior ex-
perience with a given model combined with lacking insights
into model strengths and weaknesses often lead to a certain
attachment of hydrologists to their model of choice (Addor
and Melsen, 2019). Hands-on experience with model struc-
ture uncertainty in a classroom setting, particularly through
exercises that show that the choice of model can have a
strong impact on the quality of simulations for a given catch-
ment, will prepare students to think beyond their “model of
choice”. This will prepare students for when they will need
to design modelling studies or interpret modelling results in
their future careers.

Thoughtful interpretation of model results is among many
other skills that are expected of young hydrologists (see, for
example, Table 1 in Seibert et al., 2013). However, finding
or creating course materials that cover all these expected
skills and incorporating these materials into an existing cur-
riculum is time-consuming, as is updating existing materi-
als with new knowledge. This time is consequently not spent
on preparing delivery of the material (Wagener et al., 2012).
Wagener et al. (2012) therefore introduce the Modular Cur-
rriculum for Hydrologic Advancement (MOCHA), in which
educators from many different countries freely share hydro-
logical course materials in a modular manner. Each mod-
ule addresses a specific topic and can theoretically be in-
serted into an existing curriculum with very little effort. Al-
though the MOCHA project has been inactive for some time,
the principle of freely shared, self-contained teaching mod-
ules can be of great use to the teaching community and
is experiencing a revival in platforms such as HydroLearn
(https://www.hydrolearn.org/, last access: 21 June 2022).

Seibert and Vis (2012) provide a stand-alone version of the
Hydrologiska Byråns Vattenavdelning (HBV) model that is
a good example of the MOCHA philosophy in practice. The
software is specifically modified for teaching and comes with
documentation and descriptions of various teaching goals.
This is a so-called lumped conceptual hydrological model
that relies on empirical equations to describe catchment pro-
cesses and on calibration to find its parameter values. Al-
though there is debate about the usefulness of such models
for predictions under change (e.g. see Archfield et al., 2015),
there are good reasons to use them as teaching tools provided
that the limitations of these tools are clearly communicated
to the students. Conceptual models tend to be much easier
to set up and run than their more physics-based, spatially
distributed counterparts; they generally have fewer lines of
code and internal dynamics that are easier to grasp than those
of physics-based models; and they continue to be widely
used for practical applications. These characteristics mean
that limited teaching time is spent on using and analyzing
models rather than setting up the models, that students have
more opportunity to explore internal model dynamics instead
of focusing on model outputs only and that students obtain
a firm understanding of the type of tools they are likely to
encounter in positions outside of academic research (Seibert
and Vis, 2012).

This paper introduces a set of computational exercises
designed to give students hands-on experience with model
structure uncertainty and to encourage critical thinking about
how the results of a modelling study can be interpreted. Our
goal is to increase the frequency with which model struc-
ture uncertainty is taught to (under-)graduates and to reduce
the time investment required for educators to do so. The ex-
ercises use two conceptual model structures applied to two
carefully selected catchments to illustrate various important
lessons about hydrological model selection. Briefly, these
lessons focus on the need to carefully interpret aggregated
performance metrics, the dangers of applying models in new
places based on performance elsewhere and the need to con-
sider if a model’s internal structure is an appropriate repre-
sentation of the catchment at hand.

Wagener et al. (2012) outline a need for multimedia tools
that support teaching in hydrology, specifically mentioning
“a model base with algorithms that the students can down-
load and use to support their homework assignments or
in terms projects (Wagener and et al., 2004). Such algo-
rithms need to be accompanied by sufficient documentation
and data examples. Such algorithms need to be accompa-
nied by sufficient documentation and data examples”. As
such, this module uses open-source data to allow straight-
forward application in assignments and projects. Catchment
data are obtained from the Catchment Attributes and Mete-
orology for Large-sample Studies (CAMELS; Addor et al.,
2017b), for the Middle Yegua Creek near Dime Box, Texas

Hydrol. Earth Syst. Sci., 26, 3299–3314, 2022
https://doi.org/10.5194/hess-26-3299-2022
The main goal of this teaching module is to facilitate teaching of model structure uncertainty in hydrology. Learning objectives are conveyed through comparative analysis between model results generated by students, using two conceptual model structures and two catchments and a single calibrated parameter set for each combination of model and catchment.

The models and catchments in this module are selected so that, after calibration of a single optimal parameter set per combination of model and catchment, in one of the catchments both models achieve very similar KGE scores despite the models having very different structures, while in the other catchment the models achieve very different KGE scores. This is intended to convey the following lessons to students:

1. The models have very different KGE scores in one catchment. 
   *THM:* model structure choice matters.

2. The models have very similar KGE scores in another catchment. 
   *THM:* accurate simulations are not a guarantee of hydrologic realism of the model that generated the simulations.

(ID: 8109700) and the Raging River near Fall City, Washington (ID: 12145500). Models are selected from the Modular Assessment of Rainfall-Runoff Models Toolbox (MARRMoT, Knoben et al., 2019a) and are known as “Wetland model” (ID: m02) and “Collie River basin 2 model” (ID: m03).

The goals, learning objectives and materials of this module are described in more detail in Sect. 2. The exercises are described in Sect. 3 (ready-to-use student handouts can be found in Sect. S2). Section 4 briefly summarizes the benefits of using these exercises and describes two trial applications of this module at the Technische Universität Dresden, Germany. The exercises require access to, and knowledge of, either Matlab or Octave. Course materials and source documents ready for adaptation by teachers can be downloaded through GitHub: https://github.com/wknoben/Dresden-Structure-Uncertainty (last access: 21 June 2022).

2 Description

This section describes the main teaching objectives (Sect. 2.1), the catchment data and models used (Sects. 2.2.1 and 2.2.2 respectively), an overview of provided materials, requirements and installation instructions (Sect. 2.3) and suggestions on how to integrate these exercises into an existing curriculum (Sect. 2.4).

2.1 Objectives and outline

The main goal of this teaching module is to facilitate teaching of model structure uncertainty in hydrology. Learning objectives are conveyed through comparative analysis between model results generated by students, using two conceptual
(KGE scores and a summary of these take-home messages can be seen in Fig. 1):

1. Model choice matters. Because all models are “hydrological models” it is an easy assumption to make that
the choice of model is largely one of taste or conven-
ience, rather than one of suitability for the task at hand.
Comparing the performance of both models in the Mid-
dle Yegua Creek catchment (ID: 8109700) shows that
this is not the case: the choice of model strongly affects
the accuracy of obtained simulations. In this particu-
lar case, the catchment experiences periods of no flow
which model m03 can simulate but model m02 cannot.

2. Models with very different structures can achieve vir-
tually identical efficiency scores in a given catchment.
Comparing the performance of both models in the Raging
River catchment (ID: 12145500) shows that both
achieve similar KGE scores. Logically only one (or nei-
ther) of the models can be an appropriate representa-
tion of the hydrological conditions in this catchment.
This comparison shows that achieving relatively high
efficiency scores in a given catchment is no guarantee
that the model realistically represents the dominant pro-
cesses in the catchment. In other words, high accuracy
does not necessarily mean high adequacy.

3. Reinforcing the previous point, comparing the perfor-
mance of model m03 across both catchments shows
that the model achieves higher efficiency scores than
model m02 in both places, while the catchments them-
selves are structurally very different (steep, humid and
forested, versus flat, dry and bare; catchment descrip-
tions are shown as part of the suggested exercises). Log-
ically, model m03 may be a realistic representation of
the hydrological conditions in one of the catchments,
but not both. This again shows that relatively high ef-
ciciency scores are no guarantee of the model produc-
ing “the right answers for the right reasons” (Kirchner,
2006).

4. Choosing a model based on past performance should
be done with care. Comparing the performance of
model m02 across both catchments shows that model
performance is very different in both places and that
having a “successful” model (in terms of having ob-
tained accurate simulations) for one catchment is no
guarantee that this model will produce accurate simu-
lations in a different location.

Note that this teaching module does not cover the diffi-
cult issue of defining when a given efficiency score is high
enough to consider the model as a plausible candidate for fur-
ther consideration. This requires careful use of benchmarks
(e.g. Garrick et al., 1978; Seibert, 2001; Schaefli and Gupta,
2007; Seibert et al., 2018) that dictate expectations for model
performance, which is outside the scope of this module.

2.2 Catchments and models

2.2.1 CAMELS catchment data

The CAMELS data set (Addor et al., 2017b) provides mete-
orological forcing data and a variety of catchment attributes
for 671 river basins in the contiguous USA. The catch-
ments upstream of Middle Yegua Creek near Dime Box,
Texas (USGS gauge ID: 08109700), and Raging River near
Fall City, Washington (USGS gauge ID: 12145500), are used
in this module. Our example handouts (Sect. S2; GitHub
repository) direct students to data that forms the basis of, and
expands on, the catchment descriptions given below.

Middle Yegua Creek is a water-limited catchment (aridity
fraction = 1.3, with the aridity fraction calculated as mean
annual precipitation divided by mean annual potential evap-
otranspiration) with a corresponding low runoff ratio (0.11;
mean annual runoff divided by mean annual precipitation),
low mean runoff (0.3 mm d\(^{-1}\)) and on average 30 d yr\(^{-1}\)
with no observed streamflow. Precipitation is sporadic (on average
294 d have < 1 mm precipitation) and mostly concentrated
in autumn with little to no snowfall. The catchment is rela-
tively large (615 km\(^2\)) with little variation in elevation (mean
slope = 6 m km\(^{-1}\)). Vegetation cover consists mostly of crop-
land, shrubs and low trees.

Raging River is an energy-limited catchment (aridity frac-
tion = 0.37) with a high runoff ratio (0.68), high mean runoff
(3.9 mm d\(^{-1}\)) and observed streamflow on all days in the
record. Precipitation occurs regularly (180 d with < 1 mm
precipitation) and is winter dominated, although snowfall is rare (precipitation as snow fraction = 0.04). The catch-
ment is comparatively small (80 km\(^2\)) and steep (mean
slope = 86 m km\(^{-1}\)). Vegetation cover consists nearly exclu-
sively of mixed forests.

The exercises use Daymet meteorological forcing data that
is provided as part of the CAMELS data set (Newman et al.,
2015; Addor et al., 2017b). Precipitation is part of the source
data and time series of potential evapotranspiration are esti-
imated using the Priestley–Taylor method (Priestley and Tay-
lor, 1972). The exercises use data from 1 January 1989 to
31 December 1998 for model calibration and data from 1 Jan-
uary 1999 to 31 December 2009 for evaluation.

2.2.2 MARRMoT models

The Modular Assessment of Rainfall-Runoff Models Tool-
box (MARRMoT; Knoben et al., 2019a) contains Matlab/Oc-
tive code for 46 conceptual models implemented in a sin-
gle framework. Each model requires standardized inputs and
provides standardized outputs. This means that data prepara-
tion and experiment analysis scripts have to be prepared only
once, after which running and comparing different model
structures becomes trivial. The toolbox is supported by ex-
tensive documentation, divided into the main paper describ-
This course uses MARRMoT models m02 and m03 (names refer to consistent identifiers used in all MARRMoT documentation). Both models have a single state variable and four calibration parameters but very different internal mechanics (Fig. 2). Both models require time series of precipitation \( P \) and potential evapotranspiration \( E_P \) as input. Briefly, model m02 is part of the Flex-Topo approach (Savenije, 2010) and intended to represent the dominant hydrological fluxes in a western European wetland or riparian zone. This model has a single state representing catchment storage and uses four parameters to conceptualize interception by vegetation (turning precipitation \( P \) into flux \( P_e \)), surface overland flow \( Q_{w, sof} \) from a variable contributing source area and groundwater flow \( Q_{w, gw} \). Evaporation \( E_w \) occurs at the potential rate. Model m03 is part of a study that develops a model for a semi-arid catchment in Western Australia (Jothityangkoon et al., 2001). This model also has a single state representing catchment storage but uses its four parameters to conceptualize a difference between bare soil evaporation \( E_b \) and transpiration by vegetation \( E_v \), saturation excess overland flow \( Q_{se} \) if maximum storage is exceeded and threshold-based subsurface flow \( Q_{ss} \). Full details of both models, including state equations and flux parameterizations, can be found in the Supplement of Knoben et al. (2019a).

2.3 Materials, requirements and installation

2.3.1 Provided course materials

All materials can be downloaded from https://github.com/wknoben/Dresden-Structure-Uncertainty (last access: 21 June 2022). Provided are

- example exercise handouts, including instructions to obtain and install MARRMoT (handouts are also shown in Sect. S2 for convenience);
- prepared data for the second exercise (data for the first exercise are part of the MARRMoT installation);
- an example script for model calibration for the second part of the exercise for use by educators;
- calibrated parameter sets for both models that result in the KGE scores shown in Fig. 1 for use by educators.

Note that these materials are sufficient to run the exercises with minimal effort. They do not include lecture materials to introduce the topic of model structure uncertainty. Such materials should provide a level of background and detail appropriate to the curriculum into which the exercises are inserted, which will vary between curricula.

2.3.2 Software requirements

Requirements for running MARRMoT are either Matlab with the optimization toolbox installed, or Octave. MARRMoT was developed on Matlab version 9.2.0.538062 (R2017a), with the Optimization Toolbox Version 7.6 (R2017a) and tested on Octave version 4.4.1 with the “optim” package (Knoben et al., 2019a). Note that the calibration workflow example (workflow example 4) differs slightly between Matlab and Octave 4.4.1. (See Sect. 7 in the MARRMoT user guide on Github for more details about running MARRMoT in Octave.) There are no differences in workflow example 4 between Matlab and Octave 5.2.0, thanks to a recent update to MARRMoT (Mustafa Kemal Türkeri, personal communication, 2020).

2.3.3 Installation instructions

Detailed step-by-step installation instructions for MARRMoT are included in our provided handout for Exercise 1. Briefly, download or fork and clone the MARRMoT source code at https://github.com/wknoben/MARRMoT (last access: 21 June 2022). Next, remove the folder “Octave” if Matlab will be used. Open Matlab or Octave and ensure that all MARRMoT folders are added to the Matlab/Octave path. MARRMoT is now ready to be used. Data necessary for Exercise 1 are part of the MARRMoT installation. Data necessary for Exercise 2 are part of the GitHub repository that accompanies this paper and should be distributed to the students by the teacher.

2.4 Integration in existing curriculum

Assuming the existing curriculum provides access to, and instruction in, either Matlab or Octave, integrating these exercises into the curriculum could happen along the following
lines: the exercises would be preceded by a lecture that introduces the concept of model structure uncertainty. We direct the reader to, for example, Perrin et al. (2001), Clark et al. (2011b) and Knoben et al. (2020) and the references therein for potentially useful sources with which to populate lecture materials.

Next, our two proposed exercises can be run. Broad descriptions are provided in Sect. 3 while ready-to-use student handouts are included as part of the GitHub repository and in the Supplement (Sect. S2). These exercises can be used as provided, or adapted to include more or different learning objectives. Distributing the data that underpins these exercises can either be done by referring students to the GitHub repository that accompanies this paper or by downloading the data and sharing them with the students in an alternative manner. Our example exercises include all instructions needed to obtain and install the MARRMoT source code. Students are then able to work through the exercises and will use MARRMoT to calibrate both models for both catchments, obtaining the Kling–Gupta efficiency scores shown in Fig. 1. Our proposed exercises contain guiding questions that will help students draw the correct lessons from a four-way comparison of these scores, so that they arrive at the learning objectives outlined in Sect. 2.1.

Finally, a concluding lecture can focus on (1) how other sources of uncertainty can affect performance differences between models (see Sect. 4.1 for potentially useful materials), and (2) how to effectively deal with model structure uncertainty. Approaches to deal with model structure uncertainty could, for example, be (a) designing a model from the ground up for a specific combination of catchment and study purpose rather than relying on an off-the-shelf model structure (e.g. Atkinson et al., 2002; Farmer et al., 2003; Fenicia et al., 2016), (b) quantifying model structure uncertainty through the use of model intercomparison (e.g. Perrin et al., 2001; van Esse et al., 2013; Spieler et al., 2020; Knoben et al., 2020), (c) setting more objective limits for when efficiency scores are considered acceptable by defining benchmarks that provide a context of minimum and maximum expected model performance (e.g. Schaefli and Gupta, 2007; Seibert, 2001; Seibert et al., 2018; Knoben et al., 2020), (d) attempting to define model adequacy through evaluation metrics that go beyond the use of aggregated efficiency scores that only measure accuracy and that rely on, for example, multiple metrics or data sources (e.g. Gupta et al., 2008; Kirchner, 2006; Clark et al., 2011a, b) or (e) applying model-selection or model-averaging techniques to effectively select or combine models with the appropriate strengths for a given study purpose (e.g. Neuman, 2003; Rojas et al., 2010; Schöniger et al., 2014; Höge et al., 2019).

3 Exercises

Sections 3.1 and 3.2 contain summaries of student handouts that can be used to run the two computational exercises. The student handouts themselves contain step-by-step instructions and guiding questions that take students from the comparison of KGE scores to the intended learning objectives. The full handouts can be found in Sect. S2. To facilitate modification by educators, PDFs and LaTeX source files of the student handouts that describe these exercises are also available as part of the GitHub repository.

3.1 Exercise 1: MARRMoT basics

It is recommended to first run an exercise on an individual basis that introduces students to the MARRMoT framework. In the example exercise that is provided as part of the module’s materials, students are asked to go through MARRMoT’s four workflow examples and think critically about each example and possible ways to improve it. Download and installation of the toolbox are part of the exercise. The learning objectives for this exercise are for students to

- gain a basic understanding of MARRMoT functionality;
- be able to calibrate a hydrological model and create diagnostic graphics that show the simulation results.

To achieve the learning objectives, students are asked to work through MARRMoT’s four provided workflows. Workflow example 1 shows an example of running a MARRMoT model from scratch, using a single catchment and a single parameter set. The example includes loading and preparing climatic forcing, selecting one of the MARRMoT models to use, defining the models parameters and initial states, choosing settings for the numerical solver, running the model with the specified forcing and settings, and analysis of the model simulations with the KGE objective function and qualitative plots. The predefined parameter values in this example are not well chosen and students are asked to vary the values to see how the simulations change. Manual sampling of parameter values naturally leads into workflow example 2.

Workflow example 2 replaces the arbitrarily chosen single parameter set with a random sampling procedure, using the provided parameter ranges that are part of MARRMoT. Results are visualized through qualitative plots. Students are asked to consider if a different model structure might not be more suitable than the preselected model and are directed to the MARRMoT documentation to investigate which other options are available in the toolbox. Students are asked to select a different model and rerun this workflow example, leading into workflow example 3.

Workflow example 3 shows how the code can be adapted to easily run multiple different MARRMoT models, with different numbers of parameters and state variables, from a single script. The example also includes code for visualization.
of the ensemble simulations. The example uses a randomly selected parameter set which is unlikely to give very good simulations. Students are asked to think of how to improve simulations and are asked to investigate the evaporation simulations as well as streamflow simulations.

Workflow example 4 shows an example of model calibration and evaluation and forms the basis for the model structure uncertainty exercises. Students are asked to adapt this script based on the code provided in workflow examples 1 and 3 and are asked to consider better ways to initialize model storage values.

3.2 Exercise 2: model structure uncertainty

This second exercise can be completed individually or in groups and gives the students hands-on experience with model structure uncertainty. The exercise asks students to calibrate two models (as introduced in Sect. 2.2.2) for two catchments (as introduced in Sect. 2.2.1), evaluate the resulting model simulations and think critically about the implications of their findings for model structure uncertainty. If working with groups, a possible approach would be to have each group work with a certain combination of model and catchment first, and bring the groups together for a discussion of their findings after. Groups will reach different conclusions based on which model and catchment they were assigned and a class-wide discussion is critical to impart the take-home messages of this module, because these can only be obtained by comparing the calibration and evaluation results across catchments and models (see Fig. 1). The learning objectives for this exercise are for students to

- be able to navigate model documentation and the inner workings of hydrological model code;

- critically think about the relationship between model structure, catchment structure, and model calibration and evaluation procedures, and in doing so arrive at the understanding outlined in Sect. 2.1.

As the first part of Exercise 2, students are asked to familiarize themselves with the catchments and models. Catchment data are provided in the file “Part 2 – catchment data.mat” that is part of the course materials on GitHub. Students are asked to create some exploratory figures of the meteorological data and streamflow observations, and to take a look at the catchment descriptors that are provided as part of the CAMELS data set (Addor et al., 2017b). Students are also asked to familiarize themselves with the two models by referencing the MARRMoT documentation and through an initial sensitivity analysis. Next, students are asked to calibrate both models for both catchments, using workflow example 4 as a basis for their code. This part of the exercise can take some time, partly due to the need to set up calibration and evaluation scripts and partly due to the time needed for the optimization algorithm to converge. This makes the second exercise well suited for a homework assignment or for a brief introduction to the balance between accuracy of the optimization algorithm and its convergence speed.

Finally, guiding questions in the exercise description help students to compare their calibration results in four different ways:

1. By comparing the KGE scores of both models for the Middle Yegua Creek catchment (ID: 08109700), students are expected to find that an inappropriate model choice negatively impacts the accuracy of streamflow simulations. A key characteristic of this catchment is that flow events are separated by periods during which no flow occurs. Model m03 has the ability to simulate periods of zero flow due to its threshold-based subsurface flow parameterization (see Fig. 2), whereas m02 lacks this ability and will always generate some streamflow when modelled storage exceeds zero. As a result, model m03 achieves higher efficiency scores than model m02 does.

2. By comparing the KGE scores of both models for the Raging River catchment (ID: 12145500), students are expected to understand that models with very different internal mechanics can generate equally accurate streamflow simulations (in KGE terms). This is a well-known problem in hydrological modelling (e.g. see Perrin et al., 2001; Knoben et al., 2020) and students should conclude that obtaining arbitrarily high KGE scores is no guarantee of hydrological realism.

3. By comparing the performance of model m02 across both catchments, students will see that accurate model performance (in KGE terms) in one place is no guarantee that the model will generate accurate simulations in a different catchment. This comparison warns students of the risks associated with model selection based on legacy or convenience, rather than suitability for the task at hand (Addor and Melsen, 2019).

4. By comparing the performance of model m03 across both catchments, students are expected to realize that relatively high efficiency scores do not necessarily mean that the model structure realistically represents the dominant hydrological processes in a catchment, because the model can logically do so for one catchment, but not both. This reinforces the lesson learned from the second comparison described above.

From this four-way comparison, students will gain the necessary insights to fulfill the learning objectives outlined in Sect. 2.1 and Fig. 1. Students are asked to formalize these insights into three take-home messages. A class-wide discussion can be used to reinforce this understanding.
4 Discussion

4.1 Notes on data, parameter and objective function uncertainties

This computational exercise lets students investigate the concept of model structure uncertainty in isolation, using a rather specific combination of data, models and parameter values. Our survey circulated among teachers in the hydrological sciences (see Sect. S1) indicates that model structure uncertainty is taught the least often in current curricula, whereas topics such as data uncertainty and parameter uncertainty are covered more regularly. Our suggested exercises implicitly assume that they are woven into an existing curriculum that introduces the students to other sources of modelling uncertainty and these are therefore not explicitly addressed in the provided handouts. To aid teachers in discussing the roles of other modelling-related uncertainties in the context of model structure uncertainty, we performed various brief analyses as outlined below.

4.1.1 Data uncertainty

Model calibration results are typically conditional on the data used and thus calibration outcomes can vary depending on how calibration and evaluation data are selected (e.g. see Coron et al., 2012; Fowler et al., 2016). This can be caused by differences in the statistical properties of calibration and evaluation data, leading to poor transferability of the calibrated model parameters to conditions different from those during the calibration period. More complex to quantify are measurement and estimation uncertainties that affect model input (e.g. precipitation, temperature, potential evapotranspiration estimates) and the observations the model is calibrated against (e.g. streamflow). To investigate the extent to which the findings in Fig. 1 are conditional on our somewhat arbitrary choice of calibration and evaluation periods, and hence on the idiosyncrasies of the data in either period and any statistical differences between both periods, we repeated the calibration exercise with reversed calibration and evaluation periods. Table 1 confirms that calibration and evaluation KGE scores are indeed conditional on the characteristics of the data used, but the relative performance differences remain consistent with those shown in Fig. 1. There is a reasonable difference in calibrated model performance in the Middle Yegua Creek catchment and a substantial difference in evaluation performance in this catchment, whereas both models obtain very similar performance scores in the Raging River catchment.

4.1.2 Parameter uncertainty

Calibration algorithms are constrained in their ability to identify global optimal parameter sets by their termination criteria (e.g. maximum number of iterations, objective function convergence, parameter convergence; Pechlivanidis et al., 2011) and the complexity of the parameter or objective function response surface (e.g. Duan et al., 1992). As a consequence of a rapidly varying parameter response surface, small changes in model parameter values can lead to large changes in model simulations and thus to large changes in model performance. This is particularly prevalent when the model response surface contains discontinuities, as is the case for, for example, non-smoothed threshold-based flux equations (Kavetski and Kuczera, 2007). Collectively, these issues can be referred to as parameter uncertainty. We investigate the extent to which the KGE scores shown in Fig. 1 are sensitive to these issues in two ways. First, we use a two-stage Latin hypercube sampling approach (more details can be found in the caption of Fig. 3; samples generated using the SAFE toolbox, Pianosi et al., 2015) to see if the calibration algorithm correctly identifies the general region where well-performing parameters may be found. Second, we select the 100 best-performing parameter sets in each combination of model and catchment, and compare the calibration and evaluation performance of those 100 sets to those of the calibrated values, to see whether small changes in the optimal calibrated parameter values would lead to large changes in KGE scores. Figure 3 (top four rows) shows that the calibration algorithm returns solutions in regions that match the regions of high KGE scores identified through parameter sampling. Figure 3 (bottom row) further shows that, if one were to use any of the top 100 parameter sets identified through Latin hypercube sampling, relative model performance still would be consistent with that shown in Fig. 1. There is a substantial difference in model performance in the Middle Yegua Creek catchment, whereas both models obtain very similar performance scores in the Raging River catchment.

|                | KGE calibration | evaluation |
|----------------|-----------------|------------|
| Middle Yegua Creek (08109700) | 0.48 | 0.21 |
| Raging River (12145500) | 0.90 | 0.88 |

Table 1. Kling–Gupta efficiency (KGE) scores obtained using two different definitions of calibration and evaluation periods. The columns labelled “Period 1” refer to a calibration period that ranges from 1 January 1989 to 31 December 1998 and an evaluation period that covers 1 January 1999 to 31 December 2009 (i.e. values are replicated from Fig. 1). The columns labelled “Period 2” show results obtained from calibrating the models on data from 1 January 1999 to 31 December 2009 and evaluating the models on data from 1 January 1989 to 31 December 1998.
Figure 3. Top four rows: Latin hypercube parameter samples for both models and catchments, and associated KGE scores for the calibration period. Initial sampling (light blue) covered the full parameter ranges as used during model calibration to identify regions of interest for more thorough sampling (dark blue). See Fig. S4 for a version of this figure with wider limits on the y axes. Bottom row: comparison of calibration and evaluation KGE scores of the top 100 samples (defined as highest KGE scores for the calibration period) and calibrated parameter sets.
Figure 4. Sampling uncertainty in KGE calculation, quantified as the 5th to 95th percentile of the bootstrap samples performed by the *gumboot* package. Note the different scales on the y axes.

4.1.3 Sampling uncertainty in objective function calculation

The calculation of performance metrics such as KGE can be subject to considerable sampling uncertainty (Newman et al., 2015; Lamontagne et al., 2020; Clark et al., 2021), in the sense that differences between observed and simulated values on individual time steps in a given time series can have disproportionate impacts on the overall KGE score. We quantified this uncertainty for the values presented in Fig. 1, using the *gumboot* R package developed by Clark et al. (2021). Despite considerable uncertainty in the calculated KGE values, the main learning goals are unaffected: there is a distinct difference in model performance in the Middle Yegua Creek catchment (Fig. 4a), while both models obtain similar KGE scores in the Raging River catchment (Fig. 4b).

4.1.4 Choice of objective function

It is generally accepted that the choice of objective function has a strong impact on model calibration and thus on the simulations a calibrated model produces (e.g. see Mizukami et al., 2019, and references therein). While the Kling–Gupta efficiency is derived from the components that make up the Nash–Sutcliffe efficiency (Gupta et al., 2009), KGE and NSE scores are not directly comparable and their interpretation is quite different (Knoben et al., 2019b). We therefore repeated calibration of both models for both catchments with NSE as an objective function, to assess the extent to which our choice of objective function affects the four-way comparison outlined in Fig. 1. Table 2 confirms that the choice of objective function has an impact on the actual performance scores obtained, but relative performance is consistent between KGE- and NSE-based simulations: there is a distinct difference in model performance in the Middle Yegua Creek catchment with model m03 outperforming model m02, while both models obtain similar scores in the Raging River catchment.

|                     | m02      | m03      |
|---------------------|----------|----------|
|                     | KGE      | NSE      | KGE      | NSE      |
| Middle Yegua Creek  |          |          |          |          |
| Calibration         | 0.48     | 0.29     | 0.71     | 0.44     |
| Evaluation          | 0.21     | 0.20     | 0.75     | 0.51     |
| Raging River        |          |          |          |          |
| Calibration         | 0.90     | 0.82     | 0.90     | 0.81     |
| Evaluation          | 0.88     | 0.85     | 0.90     | 0.86     |

4.2 Benefits

The main goal of the proposed exercises is to provide students with hands-on experience of the concept of model structure uncertainty. This section outlines various other benefits of using the course for both educators and students.

As mentioned, both models and catchments have been specifically selected out of much larger samples for the lessons they can convey. The exploratory work needed to do so (calibrating 40+ models for 500+ catchments) would typically be well outside of what is feasible for teaching preparation. With this selection already made, educators may spend their limited time on preparation of delivery of the course materials without having to also spend time creating the exercises. The suggested exercises expose students to a variety of different concepts that easily transfer to other disciplines and topics, such as navigating peer-reviewed literature and model documentation, and working with open-source data.
open-source software and version control through GitHub. Understanding of MARRMoT and model structure uncertainty can be leveraged into term projects or theses, providing students with a certain amount of modelling experience before their projects start. The recent publication of multiple CAMELS data sets covering the USA (Addor et al., 2017b), Chile (Alvarez-Garreton et al., 2018), Brazil (Chagas et al., 2020), Great Britain (Coxon et al., 2020) and Australia (Fowler et al., 2021) can, for example, provide the necessary data for such projects.

4.3 Trial applications at Technische Universität Dresden (Germany)

The introduced computational exercises were run and evaluated in two different settings at the Technische Universität Dresden (TU Dresden). The first application was a workshop format in June 2019 where the attendees were both students (two PhD and four MSc students) and academic or scientific staff (five). The intent of the workshop was to trial prototype exercises which could potentially be included in the curriculum of the first semester “Hydrological Modelling” module of the Hydrology Master Program at TU Dresden. The content of the “Hydrological Modelling” module is an introduction to the possibilities and restrictions of representing hydrological processes with different model types; the creation, parameterization and application of abstract models; the objective assessment of uncertainties; and a critical examination of model results. Students are introduced to basic MATLAB coding during the module and are expected to have seen and used a simple hydrological model before.

An informal evaluation was conducted during this workshop by asking the attendees to fill in a short anonymous feedback form after the course was completed. Details on the workshop setup and the feedback procedure can be found in the Supplement (Sect. S3). Attendees unanimously reported that the course was easy to follow and complete, and that the main messages were clear. Various attendees specifically noted in their answers to open questions that the exercises were helpful for better understanding the material covered during the preceding seminar on model structure uncertainty, confirming the importance of hands-on exercises to reinforce learning objectives (Thompson et al., 2012). The feedback of all attendees was used to refine and improve the exercise material.

The second application of the proposed exercises was delayed until January 2022 due to the COVID-19 pandemic. This time the exercises were offered in the intended first semester “Hydrological Modelling” module as an extracurricular exercise during the last 4 weeks of the semester. Attendance to the four 90 min sessions was thus entirely voluntary. Students were able to attend either in person or via a video conference.

The first session consisted of a lecture on model structure uncertainty, similar to the seminar given before the first workshop in 2019. During the second session, students downloaded and installed MARRMoT and individually started to work on Exercise 1. (See Sect. 3.1 consisting of MARRMoT basics on running a model, sampling parameters for a model and running several models at once.) During the third session students were introduced to model calibration with MARRMoT through the last assignment of Exercise 1 and started to explore and compare the catchment data for Exercise 2. They were asked to collect their insights by providing plots and notes on both catchments on a padlet (https://padlet.com/, last access: 22 June 2022) to foster active collaboration between the in person and online attendants. During the last session students completed Exercise 2 (see Sect. 3.2) by adapting the previously used calibration script from Exercise 1 to the new catchment data. Each student was asked to choose at least one model and one catchment, and to provide their calibration and evaluation results (KGE performance and modelled discharge) in a shared spreadsheet. This allowed the lecturer to see when students were finished and to quickly plot some aggregated results for the following classroom discussion. During the final discussion students were asked to briefly present their results and compare them with the work of their peers. All insights were discussed and condensed to collectively formulate the final results at the end of the course.

This second workshop was accompanied by an updated feedback procedure consisting of two questionnaires. One questionnaire preceded the workshop and was intended to measure students’ existing knowledge on uncertainty in hydrological modelling. The second questionnaire was handed out after the workshop and was intended to assess the changes in knowledge. Fifteen students initially showed interest in the extracurricular course and participated in the first survey, while the final course assessment at the end of the semester was answered by 10 students. There were several questions that had to be answered on a 1 to 5 scale and some open questions in both surveys. All questions were asked anonymously via a Google form and are listed in Sect. S3.

According to the results, MARRMoT proved to be an easy-to-use tool (80 % of participants found the course easy to follow/understand). All participants were able to download, install and use MARRMoT within minutes using the instructions provided in the example handouts. MARRMoT’s four workflow examples proved sufficiently documented for the students to quickly grasp the basic modelling chain (data preparation, model setup, model run and analysis of simulations) and satisfactorily complete Exercise 1. Only few Matlab-related questions had to be answered by the course supervisor. Exercise 2 required students to set up and run their own model calibration scripts. Again all students were able to adjust MARRMoT’s workflow examples with only minimal guidance and produce the expected results. Note that a script showing a possible way to complete Exercise 2 is part of the provided materials for educators and available on GitHub. Figure 5 shows that 90 % of students gained knowl-
Figure 5. Responses to the feedback form distributed after the MARRMoT exercises. Only responses to the questions that had to be answered on a 1 to 5 scale are shown; responses to open questions are summarized in Sect. S3. A value of 1 indicates disagreement and 5 indicates agreement.

Figure 6. Comparison of responses to two questions evaluating student knowledge on different uncertainty aspects before and after the MARRMoT course. A value of 1 indicates no knowledge and 5 indicates knowledge about theoretical and practical aspects of different uncertainty sources.
edge and confidence in the general area of hydrological mod-
elling as well as in the area of model structural uncertainty. Not explicitly taught areas, such as parameter and data un-
certainty, benefited as well from working through the general
modelling workflow included in the four MARRMoT work-
flow examples. Of students, 80% stated they have gained
knowledge about the difference between hydrological real-
ism and high KGE scores, and 90% agreed to have gained
knowledge on how certain model structures may be more
suitable for certain places than others. Figure 6 shows that
students improved their theoretical and practical knowledge
of uncertainty aspects in hydrological modelling by compar-
ing students’ knowledge of uncertainty before and after the
MARRMoT course.

5 Conclusions

Understanding uncertainties in the modelling process is an
important skill for graduates and necessary to interpret the
results from any modelling exercise. An informal survey cir-
culated amongst educators in the Earth sciences suggests that
model structure uncertainty is less often part of the curricu-
 lum than data and parameter uncertainty are. This paper in-
 troduces a set of ready-to-use computational exercises that
can be used to introduce the concept of model structure un-
certainty to students. Running the module requires either
Matlab or Octave. The module uses open-source hydrome-
terological data for two catchments and open-source model
code for two models, specifically selected out of a much
larger sample of catchments and models for the lessons these
pairings can convey. Students are tasked to calibrate both
models for both catchments and to evaluate the calibrated
models using data that were not used for calibration. Stu-
dents are then asked to do a four-way comparison that will
show that: (1) model choice matters, as in one of the catch-
ments both models achieve very different levels of perfor-
ance; (2) accurate model performance (in efficiency score
terms) is no guarantee of hydrological realism of the mod-
els, as in one of the catchments both models achieve very
similar levels of performance, despite having very different
internal mechanics; (3) the same applies when a single model
produces accurate simulations (in efficiency score terms) in
two different catchments, as logically the model may be re-
alistically representing one of the catchments but not both;
and (4) that the fact that a model produces accurate simu-
lations in one catchment does not guarantee that this model
will work well in another location, as the performance of one
of the models is very different in both catchments. Trial ap-
llications of this module at the Technische Universität Dresden
suggest that the module can effectively transfer these insights
in the span of two afternoons or four 90 min sessions. Data,
model code, example exercise sheets and example code to
complete the exercise are provided in a GitHub repository so
that educators wanting to teach model structure uncertainty
can focus on the delivery of these materials, rather than on
creating them.

Code and data availability. The CAMELS data set (de-
scribed in Newman et al., 2015; Addor et al., 2017b) used
in this work consists of two parts: meteorological data (https://doi.org/10.5065/D6MW2F4D; Newman et al., 2014)
and catchment attributes (https://doi.org/10.5065/D6G73C3Q;
Addor et al., 2017a). Course materials can be down-
loaded from Zenodo: https://doi.org/10.5281/zenodo.6676138
(Knoben and Spieler, 2020). The module presented in this paper has been tested with MARRMoT v1.2
(https://doi.org/10.5281/zenodo.3235664; Knoben, 2019),
MARRMoT v1.4 (https://doi.org/10.5281/zenodo.6460624; Knoben and Trotter, 2022) and MARRMoT v2.1
(https://doi.org/10.5281/zenodo.6484372; Trotter and Knoben,
2022). The DOIs can be used to find each version of the modelling
toolbox.

Supplement. The supplement related to this article is available on-
line at: https://doi.org/10.5194/hess-26-3299-2022-supplement.

Author contributions. DS conceived the idea for the workshop that
led to this publication, secured the funding and organized the work-
shop. WJMK selected the models and catchments. WJMK and
DS created the course materials and wrote the paper. DS created
the survey with help from WJMK. DS analyzed the survey results.

Competing interests. The contact author has declared that neither
they nor their co-author has any competing interests.

Disclaimer. Publisher’s note: Copernicus Publications remains
neutral with regard to jurisdictional claims in published maps and
institutional affiliations.

Acknowledgements. We gratefully acknowledge the thorough work
done by the reviewers and thank the editor for granting us the time
needed to improve this manuscript.

Financial support. This research has been supported by the
European Social Fund (grant no. 100270097).

This open-access publication was funded
by the Technische Universität Dresden (TUD).

Review statement. This paper was edited by Genevieve Ali and re-
viewed by four anonymous referees.
References

Addor, N. and Melsen, L. A.: Legacy, Rather Than Adequacy, Drives the Selection of Hydrological Models, Water Resour. Res., 55, 378–390, https://doi.org/10.1029/2018WR022958, 2019.

Addor, N., Newman, A., Mizukami, M., and Clark, M. P.: Catchment attributes for large-sample studies, UCAR/NCAR, Boulder, CO [data set], https://doi.org/10.5065/D6GG3C3Q, 2017a.

Addor, N., Newman, A. J., Mizukami, N., and Clark, M. P.: The CAMELS data set: catchment attributes and meteorology for large-sample studies, Hydrol. Earth Syst. Sci., 21, 5293–5313, https://doi.org/10.5194/hess-21-5293-2017, 2017b.

AghaKouchak, A., Nakjiri, N., and Habib, E.: An educational model for ensemble streamflow simulation and uncertainty analysis, Hydrol. Earth Syst. Sci., 17, 445–452, https://doi.org/10.5194/hess-17-445-2013, 2013.

Alvarez-Garreton, C., Mendoza, P. A., Boisier, J. P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Puelma, C., Cortes, G., Garreaud, R., McPhee, J., and Ayala, A.: The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies – Chile dataset, Hydrol. Earth Syst. Sci., 22, 5817–5846, https://doi.org/10.5194/hess-22-5817-2018, 2018.

Archfield, S. A., Clark, M., Arheimer, B., Hay, L. E., McMillan, H., Kiang, J. E., Seibert, J., Hakala, K., Bock, A., Wagener, T., Farmer, W. H., Andréassian, V., Attinger, S., Viglione, A., Knight, R., Markstrom, S., and Over, T.: Accelerating advances in continental domain hydrologic modeling, Water Resour. Res., 51, 10078–10091, https://doi.org/10.1002/2015WR017498, 2015.

Atkinson, S. E., Woods, R. A., and Sivapalan, M.: Climate and landscape controls on water balance model complexity over changing timescales, Water Resour. Res., 38, 50–1–50–17, https://doi.org/10.1029/2002WR001487, 2002.

Beven, K., Smith, P. J., and Wood, A.: On the colour and spin of epistemic error (and what we might do about it), Hydrol. Earth Syst. Sci., 15, 3123–3133, https://doi.org/10.5194/hess-15-3123-2011, 2011.

Blöschl, G. and Montanari, A.: Climate change impacts – Throwing the dice?, Hydrol. Process., 24, 374–381, https://doi.org/10.1002/hyp.7574, 2010.

Butts, M. B., Payne, J. T., Kristensen, M., and Madsen, H.: An evaluation of the impact of model structure on hydrological modeling uncertainty for streamflow simulation, J. Hydrol., 298, 242–266, https://doi.org/10.1016/j.jhydrol.2004.03.042, 2004.

Chagas, V. B. P., Chaffe, P. L. B., Addor, N., Fan, F. M., Fleischmann, A. S., Paiva, R. C. D., and Siqueira, V. A.: CAMELS-BR: hydrometeorological time series and landscape attributes for 897 catchments in Brazil, Earth Syst. Sci. Data, 12, 2075–2096, https://doi.org/10.5194/essd-12-2075-2020, 2020.

Clark, M. P., Kavetski, D., and Fenicia, F.: Pursuing the method of multiple working hypotheses for hydrological modeling, Water Resour. Res., 47, W09301, https://doi.org/10.1029/2010WR009827, 2011a.

Clark, M. P., McMillan, H. K., Collins, D. B. G., Kavetski, D., and Woods, R. A.: Hydrological field data from a modeller’s perspective: Part 2: process-based evaluation of model hypotheses, Hydrol. Process., 25, 523–543, https://doi.org/10.1002/hyp.7902, 2011b.

Clark, M. P., Vogel, R. M., Lamontagne, J. R., Mizukami, N., Kohnen, W. J. M., Tang, G., Gharari, S., Freer, J. E., Whitfield, P. H., Shook, K. R., and Papalexiou, S. M.: The Abuse of Popular Performance Metrics in Hydrologic Modeling, Water Resour. Res., 57, e2020WR029001, https://doi.org/10.1029/2020WR029001, 2021.

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

Coxon, G., Addor, N., Bloomfield, J. P., Freer, J., Fry, M., Hannonfjord, J., Howden, N. J. K., Lane, R., Lewis, M., Robinson, E. L., Wagener, T., and Woods, R.: CAMELS-GB: hydrometeorological time series and landscape attributes for 671 catchments in Great Britain, Earth Syst. Sci. Data, 12, 2459–2483, https://doi.org/10.5194/essd-12-2459-2020, 2020.

Di Baldassarre, G. and Montanari, A.: Uncertainty in river discharge observations: a quantitative analysis, Hydrol. Earth Syst. Sci., 13, 913–921, https://doi.org/10.5194/hess-13-913-2009, 2009.

Duan, Q., Sorooshian, S., and Gupta, V.: Effective and efficient global optimization for conceptual rainfall-runoff models, Water Resour. Res., 28, 1015–1031, https://doi.org/10.1029/91WR02985, 1992.

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

Farmer, D., Sivapalan, M., and Jothishyangkoon, C.: Climate, soil, and vegetation controls upon the variability of water balance in temperate and semiarid landscapes: Downward approach to water balance analysis, Water Resour. Res., 39, 1035, https://doi.org/10.1029/2001WR000328, 2003.

Fenicia, F., Kavetski, D., Savijeni, H. H. G., and Pfister, L.: From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions, Water Resour. Res., 52, 954–989, https://doi.org/10.1002/2015WR017398, 2016.

Fowler, K. J. A., Peel, M. C., Western, A. W., Zhang, L., and Peterson, T. J.: Simulating runoff under changing climatic conditions: Revisiting an apparent deficiency of conceptual rainfall-runoff models, Water Resour. Res., 52, 1820–1846, https://doi.org/10.1002/2015WR018068, 2016.

Fowler, K. J. A., Acharya, S. C., Addor, N., Chou, C., and Peel, M. C.: CAMELS-AUS: hydrometeorological time series and landscape attributes for 222 catchments in Australia, Earth Syst. Sci. Data, 13, 3847–3867, https://doi.org/10.5194/essd-13-3847-2021, 2021.

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

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling the second and third workshops, J. Hydrol., 320, 3–17, https://doi.org/10.1016/j.jhydrol.2005.07.031, 2006.

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling the second and third workshops, J. Hydrol., 320, 3–17, https://doi.org/10.1016/j.jhydrol.2005.07.031, 2006.

Gupta, H. V., Wagener, T., and Liu, Y.: Reconciling the second and third workshops, J. Hydrol., 320, 3–17, https://doi.org/10.1016/j.jhydrol.2005.07.031, 2006.
of calibration metrics for “high-flow” estimation using hydrologic models, Hydrol. Earth Syst. Sci., 23, 2601–2614, https://doi.org/10.5194/hess-23-2601-2019, 2019.

Nash, J. and Sutcliffe, J.: River flow forecasting through conceptual models part I – A discussion of principles, J. Hydrol., 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Neuman, S. P.: Maximum likelihood Bayesian averaging of uncertain model predictions, Stoch. Environ. Res. Risk A., 17, 291–305, https://doi.org/10.1007/s00477-003-0151-7, 2003.

Newman, A., Sampson, K., Clark, M. P., Bock, A., Viger, R. J., and Blodgett, D.: A large-sample watershed-scale hydrometeorological dataset for the contiguous USA, UCAR/NCAR, Boulder, CO [data set], https://doi.org/10.5065/D6MW2F4D, 2014.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q.: 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, Hydrol. Earth Syst. Sci., 19, 209–223, https://doi.org/10.5194/hess-19-209-2015, 2015.

Peclivanidis, I., Jackson, B., McIntyre, N., and Wheeler, H.: Catchment-scale hydrological modelling: a review of model types, calibration approaches and uncertainty analysis methods in the context of recent developments in technology and applications, Global NEST, 13, 193–214, 2011.

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

Pianosi, F., Sarrazin, F., and Wagener, T.: A Matlab toolbox for Global Sensitivity Analysis, Environ. Model. Softw., 70, 80–85, https://doi.org/10.1016/j.envsoft.2015.04.009, 2015.

Priestley, C. H. B. and Taylor, R. J.: On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters, Mon. Weather Rev., 100, 81–92, https://doi.org/10.1175/1520-0493(1972)100<0081:OTAOSH>2.3.CO;2, 1972.

Rojas, R., Kahunde, S., Peeters, L., Batelaan, O., Feyen, L., and Dassargues, A.: Application of a multimodel approach to account for conceptual model and scenario uncertainties in groundwater modelling, J. Hydrol., 394, 416–435, https://doi.org/10.1016/j.jhydrol.2010.09.016, 2010.

Savenije, H. H. G.: Topography driven conceptual modelling (FLEX-Topo), Hydrol. Earth Syst. Sci., 14, 2681–2692, https://doi.org/10.5194/hess-14-2681-2010, 2010.

Schaefl, B. and Gupta, H. V.: Do Nash values have value?, Hydrol. Process., 21, 2075–2080, https://doi.org/10.1002/hyp.6825, 2007.

Schöninger, A., Wöhlle, T., Samaniego, L., and Nowak, W.: Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence, Water Resour. Res., 50, 9484–9513, https://doi.org/10.1002/2014WR016062, 2014.

Seibert, J.: On the need for benchmarks in hydrological modelling, Hydrol. Process., 15, 1063–1064, https://doi.org/10.1002/hyp.446, 2001.

Seibert, J. and Vis, M. J.: Teaching hydrological modelling with a user-friendly catchment-runoff-model software package, Hydrol.
Earth Syst. Sci., 16, 3315–3325, https://doi.org/10.5194/hess-16-3315-2012, 2012.
Seibert, J., Uhlenbrook, S., and Wagener, T.: Preface Hydrology education in a changing world, Hydrol. Earth Syst. Sci., 17, 1393–1399, https://doi.org/10.5194/hess-17-1393-2013, 2013.
Seibert, J., Vis, M. J. P., Lewis, E., and van Meerveld, H.: Upper and lower benchmarks in hydrological modelling, Hydrol. Process., 32, 1120–1125, https://doi.org/10.1002/hyp.11476, 2018.
Spieler, D., Mai, J., Craig, J. R., Tolson, B. A., and Schütze, N.: Automatic Model Structure Identification for Conceptual Hydrologic Models, Water Resour. Res., 56, e2019WR027009, https://doi.org/10.1029/2019WR027009, 2020.
Thompson, S. E., Ngambeki, I., Troch, P. A., Sivapalan, M., and Evangelou, D.: Incorporating student-centered approaches into catchment hydrology teaching: A review and synthesis, Hydrol. Earth Syst. Sci., 16, 3263–3278, https://doi.org/10.5194/hess-16-3263-2012, 2012.
Trotter, L. and Knoben, W. J. M.: MARRMoT v2.1 (v2.1), Zenodo [code], https://doi.org/10.5281/zenodo.6484372, 2022.

van Esse, W. R., Perrin, C., Booij, M. J., Augustijn, D. C. M., Fenicia, F., Kavetski, D., and Lobligois, F.: The influence of conceptual model structure on model performance: a comparative study for 237 French catchments, Hydrol. Earth Syst. Sci., 17, 4227–4239, https://doi.org/10.5194/hess-17-4227-2013, 2013.
Wagener, T. and McIntyre, N.: Tools for teaching hydrological and environmental modeling, Comput. Educ. J., 17, 16–26, 2007.
Wagener, T., Gupta, H. V., Carpenter, K., James, B., Vazquez, R., Sorooshian, S., and Shuttleworth, J.: A hydroarchive for the free exchange of hydrological software, Hydrol. Process., 18, 389–391, https://doi.org/10.1002/hyp.5216, 2004.
Wagener, T., Kelleher, C., Weiler, M., McGlynn, B., Gooseff, M., Marshall, L., Meixner, T., McGuire, K., Gregg, S., Sharma, P., and Zappe, S.: It takes a community to raise a hydrologist: The Modular Curriculum for Hydrologic Advancement (MOCHA), Hydrol. Earth Syst. Sci., 16, 3405–3418, https://doi.org/10.5194/hess-16-3405-2012, 2012.