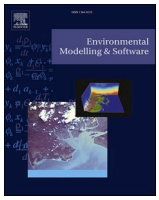
Environmental Modelling and Software 167 (2023) 105779

Contents lists available at ScienceDirect

Environmental Modelling and Software

journal homepage: www.elsevier.com/locate/envsoft



On how data are partitioned in model development and evaluation: Confronting the elephant in the room to enhance model generalization☆

Holger R. Maier [a,*], Feifei Zheng [b], Hoshin Gupta [c], Junyi Chen [b], Juliane Mai [d,e], Dragan Savic [f,g], Ralf Loritz [h], Wenyan Wu [i], Danlu Guo [j], Andrew Bennett [c], Anthony Jakeman [k], Saman Razavi [l], Jianshi Zhao [m]

[a] School of Architecture and Civil Engineering, The University of Adelaide, Adelaide, Australia
[b] College of Civil Engineering and Architecture, Zhejiang University, China
[c] Department of Hydrology and Atmospheric Sciences, The University of Arizona, Tucson, USA
[d] Computational Hydrosystems, Helmholtz Centre for Environmental Research–UFZ, Leipzig, Germany
[e] Center for Scalable Data Analytics and Artificial Intelligence–ScaDS.AI, Leipzig, Germany
[f] KWR Water Research Institute, the Netherlands
[g] Centre for Water Systems, University of Exeter, Exeter, United Kingdom
[h] Karlsruhe Institute of Technology (KIT), Institute of Water and River Basin Management, Karlsruhe, Germany
[i] Department of Infrastructure Engineering, The University of Melbourne, Melbourne, Australia
[j] School of Engineering, College of Engineering, Computing & Cybernetics, Australian National University, Canberra, Australia
[k] Insitute for Water Futures, Fenner School of Environment and Society, Australian National University, Canberra, Australia
[l] Institute for Water Futures, Mathematical Sciences Institute, Australian National University, Canberra, Australia
[m] Department of Hydraulic Engineering, Tsinghua University, Beijing, China

ARTICLE INFO

Handling editor: Daniel P Ames

Keywords:
Model development
Model evaluation
Data partitioning
Data splitting
Calibration
Validation
Uncertainty
Earth systems

ABSTRACT

Models play a pivotal role in advancing our understanding of Earth's physical nature and environmental systems, aiding in their efficient planning and management. The accuracy and reliability of these models heavily rely on data, which are generally partitioned into subsets for model development and evaluation. Surprisingly, *how* this partitioning is done is often not justified, even though it determines what model we end up with, how we assess its performance and what decisions we make based on the resulting model outputs. In this study, we shed light on the paramount importance of meticulously considering data partitioning in the model development and evaluation process, and its significant impact on model generalization. We identify flaws in existing data-splitting approaches and propose a forward-looking strategy to effectively confront the "elephant in the room", leading to improved model generalization capabilities.

1. Introduction

Computer-based models are used extensively to help better understand Earth and environmental problems, and to manage various environmental and water resource systems. The use of data is central to the development of these models, as well as their practical applications (Vilas et al., 2023). Unfortunately, the choice of *which* data are available for model development is often beyond the control of modelers (Gibbs et al., 2018; Li et al., 2012). However, modelers do have a choice about *how* the available data are used. This generally relates to the manner in which the available data are partitioned into model *development* and *evaluation* subsets (Yang et al., 2018; Mai, 2023) - a process that is commonly referred to as "*data-splitting*" (e.g., Picard and Berk, 1990; Arsenault et al., 2018; Liu et al., 2018).

Typically, the first portion of the available data is used for model *development* (e.g. 60% or 70%), while the remaining portion (e.g. 40% or 30%) is used for model *evaluation* (Addor and Melsen, 2019; Mount et al., 2016). In some instances, the model *development* data are split further into *selection* and *calibration* subsets, where the *selection* subset is used for model structure determination and/or its hyperparameter

☆ Members of Editorial Board for this journal: Anthony Jakeman, Holger Maier, Saman Razavi, Wenyan Wu.

* Corresponding author.
E-mail address: holger.maier@adelaide.edu.au (H.R. Maier).

https://doi.org/10.1016/j.envsoft.2023.105779
Received 5 June 2023; Received in revised form 23 July 2023; Accepted 27 July 2023
Available online 31 July 2023

tuning, and the *calibration* subset for tuning the parameters of a model with a given structure. In contrast, the *evaluation* subset is used for independent validation/testing of the generalization performance of the developed model (Coron et al., 2012; Wu et al., 2014; Maier et al., 2023). However, this additional data splitting step is not required if a single model structure is considered.

An alternative but related approach to the above data-splitting technique is the use of cross-validation, where models are calibrated and evaluated on multiple subsets of the data (see pioneering work on splines by Craven and Wahba, 1978). Three examples are: calibrating on subsets involving a moving window of the data and evaluating on the remainder (e.g., Coron et al., 2012); calibrating in multifarious ways on a majority of the data but omitting/holding out a much smaller subsample each time, which is useful when confronted with limited amounts of data; or calibrating on different conditions in the data such as those related to climate, for instance seasonal forcings and inter-annual variabilities (e.g., Trotter et al., 2023).

The primary advantage of cross-validation is that a richer set of results is obtained for which model performance can be assessed, or at least detected, according to its variability in both calibration and evaluation, thereby providing a pragmatic appreciation of model uncertainty. In addition, cross-validation can provide an indication of the conditions under which a model performs to varying degrees, as well as highlight its limitations. In yielding a number of parameter sets corresponding to the different calibration splits, the parameter sets can be used to generate a combined or ensemble model; alternatively, the model parameterization that produces the most robust model, based on evaluation performance, can be chosen.

Although papers reporting on modeling studies generally provide details on ***how*** the available data are divided into their respective subsets, very few justify ***why*** the selected data split has been selected. This is despite the significant body of literature pointing to the importance of attending to the characteristics of the data used for model development and evaluation, such as informativeness (Gupta et al., 1998; Vrugt et al., 2003; Singh and Bárdossy, 2012) - an aspect that is commonly accepted in statistical learning (Hastie et al., 2009). The consequences of ignoring the impacts of ***how*** the available data are split into their respective subsets can be far reaching, potentially resulting in models with reduced generalization ability, a misleading assessment of the absolute and relative performance of models, and suboptimal decisions based on the outputs of these models (Chen et al., 2022; Zheng et al., 2022). These consequences arise because whichever data are used for model development and evaluation has a direct impact on which model is actually developed (e.g., values of the calibrated model parameters), as well as its perceived performance (Bowden et al., 2002; May et al., 2010).

In order to confront the elephant in the room,[1] this position paper (i) articulates why a careful choice of ***how*** the data are partitioned into development and evaluation subsets is critical, (ii) outlines ***problems*** with both traditional and state-of-the-art data-splitting approaches, and (iii) provides a ***roadmap and call to action***. We anticipate that this paper will stimulate discussion on what role data play in the model development process, what the purpose of model calibration and validation actually is, and how we should evaluate the performance of models. We also believe that such attention can only result in the development of improved approaches to data-splitting, and hence more credible models.

2. The importance of data splitting – the elephant in the room

To emphasize the importance of proper data splitting, it is helpful to consider the analogy to a student taking a hydrology course (Table 1). In this context, the role of teacher (model developer) is to both facilitate optimal student learning (model development) and to evaluate how well the student has learned the relevant material (model evaluation). To facilitate learning, we provide students with in-depth exposure to, and opportunities to engage with, relevant course topics, such as providing homework questions for refining their understanding. Then, to ascertain the degree of student competency, we test student performance by designing exam questions on some, if not all, of the same topics, using questions that are generally conceptually similar, but not the same, as those previously assigned as homework. If the exam questions were exactly the same as the homework questions, it would be impossible to tell whether students had internalized the information sufficiently well to be able to successfully answer new (but related) questions, or whether they had simply memorized the solutions.

Table 1
Details of analogies between teaching a hydrology course and developing a model.

Hydrology Course Teaching	Model Development
Teacher	Modeler
Student understanding	Learned model representation
Course topics to be covered	Types of events to be modeled
Homework questions	Conditions represented in model development data
Exam questions	Conditions represented in model evaluation data

Analogously, as part of the model development process, our goal is for models to internalize an appropriate representation of the system's structure (i.e., "learn" appropriate input-output mappings) so that they can effectively make predictions under the diverse boundary conditions and modeling objectives they were designed for (e.g., high, medium, and low flow conditions, wet and dry prior climatic conditions, warm and cool temperatures, etc. – akin to the variety of topics covered in the hydrology course). For a given model structure, this is achieved by adjusting the model parameters to minimize the difference between modeled and corresponding measured outputs for a set of example input-output data points (in which sequential/spatial patterns are analogous to homework questions given to students). In addition, we would also like to evaluate (validate) the performance of the calibrated models under a breadth of similar conditions (course topics), but using different input-output data points (exam questions).

The reason we need to use different data for model development and evaluation is that, like students, models are potentially able to "*memorize*" input-output mappings for particular patterns of data points without properly internalizing (learning) the underlying representations (principles, processes, and relationships). In model development, this is referred to as "*overfitting/overtraining*" (Vaze et al., 2010; Lever et al., 2016; Razavi, 2021). Clearly, to properly achieve the above goals, both types of data subsets (i.e., the ones used for model development and model evaluation) must comprehensively cover the full range of relevant hydro-climatic conditions the model is intended to be representative of, or we run the risk of biasing the model towards performing well under one limited set of conditions at the expense of some others. By analogy, if all homework questions are focused on one topic within the course, then students are likely to understand this topic at the expense of understanding the other topics (Wu et al., 2013).

The above examples highlight three key model-development principles.

1) **It is imperative that data sets used for model development and evaluation are different.** To enable the true generalization ability of models (knowledge of students) to be tested, rather than how well they have memorized the input-output relationships in the model development data (homework questions), it is important that different data (questions) are used for model development (homework) and evaluation (exam) (Maier et al., 2010; Humphrey et al., 2017). While teachers can use their experience and understanding to generate a wide variety of representative homework and exam questions, modelers typically only have access to a fixed data set,

[1] "An elephant in the room" is an English expression that means that there is an obvious problem that people do not want to talk about.

which therefore needs to be split into development and evaluation subsets. Of course, it is then a simple example of a cross-validation to reverse the process by calibrating on the evaluation set and evaluating on the calibration set, thereby gaining an enhanced appreciation of model performance. In more intricate cases, the number of development and evaluation sets may be several, and where data are limited there may be some overlap between data sets.

2) **It is vital that all types of patterns/events contained in the available data that are relevant to the modeling purpose are included in the model development subset (i.e. for selection and calibration).** To ensure that models (students) have the opportunity to develop the best possible representation of the system (best understanding of the course topics), the model development data (homework questions) must cover all types of events that are represented in the available data (all course topics). Without this, neither the generalization potential of the model, nor the learning potential of the student, can be maximized.
3) **It is ideal that all types of patterns/events contained in the available data are *also* included in the model evaluation subset.** To assess the potential generalization ability of the model (level of competence achieved by the student), the model evaluation data (exam questions) must seek to cover all types of events that are represented in the available data (all course topics). Without this, the potential ability of the calibrated model to perform well in practice remains poorly understood, while the understanding of the student remains untested, both of which are undesirable as they present an incomplete and potentially misleading picture of future performance. This is especially important to consider when the resulting model will be used for modeling of specific event types (such as peak flows) to ensure that the model is actually fit for its intended purpose (e.g., see Hamilton et al. (2022) on characterizing fitness for purpose in terms of usability, reliability and feasibility).

In general, the latter two principles can be satisfied relatively easily for the hydrology course example, since the teacher simply needs to generate (or select) an appropriate number of homework and exam questions from each of the course topics. However, for model development and evaluation, given that the available data are fixed, the modeler must deal with the facts that (i) the types of events covered by the data might be poorly informative/representative, and (ii) the number of data points (or data patterns) corresponding to different types of events are often highly variable – in general, infrequently experienced events (such as extreme events, which might be highly important to be able to properly simulate) constitute only small fractions of the data and are therefore represented in an unbalanced manner. This can make it difficult, and sometimes impossible, to ensure that sufficient examples of the different types of events are properly included in both the model development and evaluation subsets, making it extremely challenging to ensure that the second and third principles are satisfied. Consequently, great care needs to be taken when assigning the available data to development and evaluation subsets to avoid perverse and/or misleading outcomes. This issue is exacerbated when the model development data need to be split further into selection and calibration subsets.

When speaking about generalization ability, we must point out the need to distinguish between the ability of models to generalize well under "*in-sample*" conditions – in the sense of learning a representation that enables it to perform well under conditions that are well represented by the entire available data sample, as opposed to its ability to generalize to "*out-of-sample*" conditions – in the sense of learning a representation that enables it to perform well in extrapolation, that is under conditions that have not yet been seen (e.g., see discussion in Klemes (1986), Razavi (2021), Shen et al. (2022), as well as literature on "*prediction in ungauged basins*" (Hrachowitz et al., 2013)). We can think of the former as an "*Order-one*" generalization ability, and the latter as an "*Order-two*" generalization ability. Out-of-sample generalization is clearly a more challenging problem, because it requires the model development process to be able to detect, extract and construct representations of deeper underlying process principles that might not be apparent from a more superficial "*pattern-matching*" (e.g., loss-function minimization) based analysis of the available data; it may even require access to a much larger database (Gupta et al., 2014; Gauch et al., 2021) from which such deeper underlying process principles can be extracted and learned (Kratzert et al., 2018; Lees et al., 2022). The important point is that "*Order-one*" generalization ability can be thought of as a necessary prerequisite to achieving "*Order-two*" generalization ability – if the model is unable to perform robustly under the conditions for which it has been trained, we should be skeptical about its ability to generalize to new, unseen conditions.

By analogy, if, as part of their exam, students are tested on their ability to answer questions that require insights into hydrological science deeper than have been covered by the training components of the course (lectures, discussions, and homework questions), then we can reasonably first require that they perform well on "*in-sample*" examination questions (i.e., the pre-requisites to deeper understanding), before going on to test their ability to generalize (i.e., extrapolate) *beyond* the content covered in the course. In the discussion that follows, we focus primarily on how to ensure the ability of models to achieve "*in-sample*" (Order-one) generalization ability.

2.1. The problem with traditional data-splitting approaches

As mentioned above, the traditional and most common approach to data-splitting is to use the first (or sometimes last) X% (say 60%) of the available time series of data for model development (model selection and parameter calibration) and the remainder for evaluation (validation/testing). Although not generally mentioned explicitly, the most likely rationale for this approach is that it mimics what happens in practice once a model has been deployed (Shen et al., 2022; Mai, 2023). However, this approach has the potential to result in a number of perverse outcomes.

Most important of these is, that by using the first X% of the available data for model development and the remainder for evaluation, the modeler has no oversight over which types of events are included in either data set. For example, the first part of the available data used to develop a rainfall-runoff model may consist mainly of low-flow events while the second part may consist mainly of high-flow events. Per our hydrology course analogy, this would be analogous to covering, for example, only groundwater-related material during the course and then assessing the ability of students to address questions related to surface water hydrology during the exam. Clearly, not only can this undermine the ability to achieve the best possible outcomes given available information, but it is also likely to result in a misleading assessment of real performance.

Specifically, the potential for generalization is limited to aspects of groundwater hydrological understanding (covered in the course) that translate in some meaningful manner to the surface water realm. Further, not covering surface water hydrology during the training phase would undermine the ability of the student to become a competent hydrologist. Similarly, by not presenting examples of high-flow events during model development, we undermine the ability of the model to perform well over the widest range of events possible, and by evaluating the model on events it has not seen during model development we would obtain a misleading assessment of the quality of the model development process.

The degree to which this data-splitting problem occurs will largely depend on the temporal (or spatial) distribution of events in the data (Zheng et al., 2018; Guo et al., 2020; Chen et al., 2022). Taking catchment rainfall-runoff modeling again as an example, if the distribution of different types of hydrological events is relatively even over time (see Fig. 1a), then the above issue is less likely to occur, due to all kinds of events being equally likely to be included in both the model

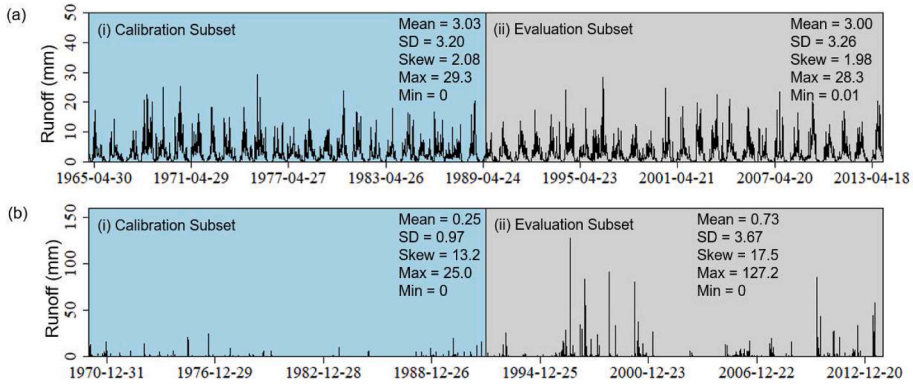


Fig. 1. Example runoff time series data from two Australian catchments (Raupach et al., 2009; 2012), showing (a) a time series where the different types of events that are contained in the data are relatively evenly distributed over time, resulting in similar statistical properties of the model development and evaluation subsets and (b) a time series where the different types of events that are contained in the data are not evenly distributed over time, resulting in dissimilar statistical properties of the development and evaluation subsets. SD: standard deviation of the runoff data, Skew: skewness of the runoff data.

development and evaluation subsets, as evidenced by the similarity in the statistical properties of the two subsets in the example data in Fig. 1a. However, if the different types of events are not evenly distributed over time (Fig. 1b), the types of events that are included in the development and evaluation data can be very different, as evidenced by the dissimilarity in the statistical properties of the two subsets in the example data in Fig. 1b. Consequently, when traditional approaches to data splitting are used, whether a model can generalize well or not is *completely arbitrary*, as it is contingent on the distribution of the data, over which modelers have no control. In addition, the model developer has no idea which of these two cases applies. This means that there is a large "random" element to both the resulting model (i.e., which values the calibrated parameters take) and the evaluation performance of the model.

2.2. The problem with state-of-the-art data-splitting approaches

To provide the best opportunity for the development process to result in models that can generalize well over a variety of different types of events, whilst also providing a rigorous independent assessment of model generalization ability, the method for partitioning the data should seek to ensure that the stochastic properties of, and the processes embedded in, the development and evaluation subsets are as similar as possible. A number of data-splitting methods have been developed that try to achieve this goal by using a variety of approaches. Examples include.

1) Using theoretical hydrological understanding of system processes and behaviors to pre-label different kinds of events in the data and then allocating these to the different subsets (Seibert and McDonnell, 2015).
2) Using formal optimization approaches to allocate individual data points to the respective subsets so as to minimize the difference between key statistical properties (e.g. mean, standard deviation, skewness, etc.) of the different data subsets (e.g. Bowden et al., 2002).
3) Using deterministic allocation processes that ensure that the data points allocated to the different subsets correspond to similar "patterns" in the data based on their relative degree of closeness to each other using an appropriate distance metric (e.g. DUPLEX (Snee, 1977; May et al., 2010), Modified DUPLEX (MDUPLEX) (Chen et al., 2022)).
4) Using clustering approaches to partition the available data into similar regions that correspond to the different types of events that can be detected in the data (see Bowden et al., 2002). Samples from each region are then allocated to the development and evaluation subsets in accordance with some desired percentage split using an appropriate sampling approach (e.g., self-organizing map (SOM) based proportional sampling (SBSS–P) (May et al., 2010), coupled SOM-DUPLEX (SOMPLEX) (Chen et al., 2022)).
5) Using k-fold cross validation, where multiple splits are used to perform multiple calibrations (specifically k splits, where k is some integer number) and subsequent validations. This does raise some questions about how to combine the multiple calibration parameters, but approaches like Bayesian model averaging (Duan et al., 2007; Wöhling et al., 2015) can address this question. Additionally, the use of k-fold cross validation uses all of the available data and can be used to compare evaluation in each of the splits, which can provide insight into performance for particular events such as extreme rainfall, and in more general terms into what is the most robust model parameterization.

While such state-of-the-art data-splitting methods are better at ensuring mutual consistency of the statistical properties of the model development and evaluation data subsets than traditionally used strategies, they face two primary challenges.

1) **Extreme values in the available data generally result in the bias-variance dilemma.** When deterministic data-splitting methods are used (e.g., DUPLEX (Snee, 1977; May et al., 2010) and MDUPLEX (Chen et al., 2022)), a single data split is obtained. Consequently, extreme values have to be either allocated to the development or evaluation subset, most likely resulting in some degree of dissimilarity (i.e., bias) between the statistical properties of the two data subsets, which is undesirable. However, given that these approaches are deterministic, the same data split is obtained every time they are applied (i.e., zero variance), which is desirable. In contrast, when stochastic data-splitting methods are used (e.g., SBSS-P (May et al., 2010), SOMPLEX (Chen et al., 2022)), a number of data splits are obtained. Consequently, extreme values can be allocated to either development or evaluation subsets in different splits, enabling the average difference between the statistical properties of the subsets (i.e., bias) to be reduced, which is desirable. However, as different splits are obtained every time the method is applied, there is some variance, which is undesirable. Consequently, there are bias-variance trade-offs between the different state-of-the-art methods. Although some guidance is available on their application based on the properties of the available data (e.g., Wu et al., 2013; Zheng et al., 2018; Guo et al., 2020; Chen et al., 2022), a decision regarding these trade-offs needs to be made by modelers when selecting a state-of-the-art data-splitting approach.
2) **Their application to "path-dependent" models that possess memory representations requires additional care.** Preserving the time-order, or path-dependency, of data is critical to all process-based models and some data-driven models that are memory-enabled, such as recurrent neural networks (RNNs) and their popular variation, long short-term memory (LSTM) networks (Maier et al., 2023). This is because the input-state-output trajectories generated by these models can be highly sensitive to the values used to initialize the system states and more generally to the system states in the

previous time steps (e.g., the initial level of soil water storage). In other words, their states are dependent on system components having long residence times (memory) that persist for extended periods (Hoell et al., 2017). As a practical consequence, such models cannot (in general) be developed via a parameter calibration process that is based on the use of randomized, mini-batches of data, as is commonly adopted in the development of many data-driven models (Kratzert et al., 2018). Therefore, to ensure that the model states evolve in a manner that is consistent with the dynamics of the physical system, we are restricted to the use of calibration methods wherein the data fed to the model during calibration must consist of temporally-consecutive inputs. This restricts the application of data-splitting methods that are designed to assign data without maintaining their underlying time structure and is likely the reason why state-of-the-art data-splitting approaches have been applied primarily in the data-driven modeling domain (e.g., Bowden et al., 2002; May et al., 2010; Wu et al., 2013; Zheng et al., 2018; Guo et al., 2020; Chen et al., 2022).

3. A roadmap of the way forward – Confronting the elephant in the room

When it comes to developing Earth and environmental systems models, the impact of data splitting is undoubtedly the elephant in the room (Fig. 2). Even though data splitting occurs near the beginning of the model development process, and therefore impacts *all* subsequent processes, its impact is largely ignored, with little to no justification of why the available data are split the way they are and seemingly no understanding of, or interest in, the potential impacts of these choices. This is despite the fact that the way the data are split determines what model we end up with, how we assess its performance, the resulting uncertainty in model outputs and, ultimately, what decisions we make based on these model outputs (Fig. 2). This is extremely surprising, given the large amount of attention that is paid to other model development processes, such as calibration (van Vliet et al., 2016), evaluation (Bennett et al., 2013) and uncertainty/sensitivity analysis (Ascough et al., 2008; Pianosi et al., 2016; Razavi et al., 2021; Saltelli et al., 2021), as well as the significantly greater degree of scrutiny that these processes are subjected to, even though the outcomes of these processes are *all* contingent on how the available data are split into their respective subsets. This is somewhat analogous to focusing on the noise, while ignoring the signal. Conversely, if data splitting is undertaken in a more considered manner, these methods of calibration, sensitivity and uncertainty analysis etc. can be applied in a cross-validation setting, thereby strengthening their value.

In order to confront the elephant in the room and improve modeling practice, we call on the modeling and research communities to focus on the following courses of action (Fig. 3).

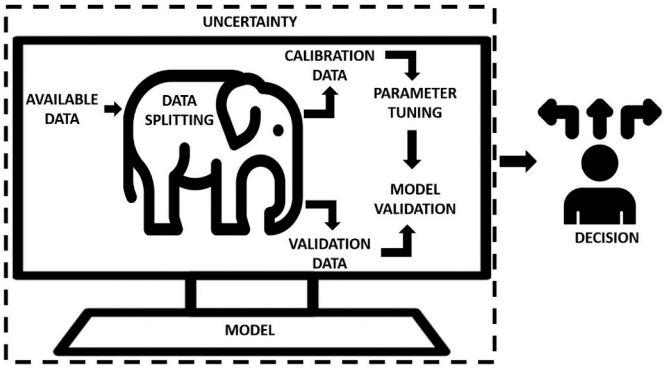


Fig. 2. Impacts of data splitting on a number of stages in the model development process and consequently on the model outputs and subsequent decisions made based on these outputs. These impacts have been virtually ignored by the modeling community, making it the "elephant in the room".

1) **Increase Awareness and Adoption.** Researchers and modelers need to become more aware of the widespread impacts the way the available data are split into development and evaluation subsets has on the resulting model and its perceived performance. This is the case for all domains within the Earth and environmental sciences, especially due to the prevalence of data that are highly temporally variable and that contain extreme events (e.g., floods, droughts, typhoons and heat waves), which exacerbates the impacts of a lack of attention to the impacts of data splitting. Consequently, modelers and researchers alike need to act to adopt appropriate data-splitting practices as part of routine model development. Here, editors and reviewers of modeling journals can play an important role in demanding that greater attention be given to the justification of how the available data are used for model development and evaluation and report on any implications of the choices made. The modeling community should also promote the need to appropriately address the data-splitting issue when developing guidelines for modeling best practices. This assists in assessing model relevance, especially the range of applicability of models. In some sense this is a holy grail problem when dealing with Earth system models or any of their components driven by climate dynamics. It is possible we will never have a model that works well under all historical and future conditions, making the above recommendation for assessing applicability very important.
2) **Perform Comparison and Evaluation.** Researchers and modelers need to quantify the impact of data splitting (both what fraction(s) of the available data are used for model development and evaluation and how the available date are apportioned to each of these), relative to that of other steps in the model development process (e.g., Jakeman et al., 2006). This is because, in addition to the way the data are split into model development and evaluation subsets, model errors or biases can also be induced at other stages of the model development processes, such as the selection of model structure and type, the choice of calibration algorithm and even in the selection of objective functions. However, the relative influence of these choices on these errors and biases will be unknown unless they are investigated in context. Consequently, a focus of future research efforts should be developing procedures that distinguish errors emanating from different stages of the modeling process and how they relate to the ways data splitting is undertaken. In addition, there can be much usefulness in evaluating alternative approaches to identifying appropriate data-splitting practices for different study areas, models and their accompanying data. At the very least, modelers should test a range of data-splitting options and associated cross-validation exercises for a particular case study application to gain an improved understanding of model performance and its limitations for the problem at hand.
3) **Develop Improved Data-Splitting Approaches.** Research is required to improve a number of aspects of data-splitting approaches, including:
   a. **Addressing the bias-variance dilemma.** Although some efforts have been made to develop data-splitting methods that minimize the bias-variance dilemma for data with different characteristics (e.g., May et al., 2010; Zheng et al., 2018; Chen et al., 2022), further work is required to ensure the "best" possible models, given the available data, are developed. This is particularly the case when dealing with noisy data, extreme events and data with underlying trends and patterns that are associated with changing forcing conditions (such as climate). A promising avenue for development is the adoption of a two-step calibration process, where the first step requires data splitting to obtain insights on model transferability and to quantify uncertainties (e.g., by using k-fold cross-validation), while the second step uses all available

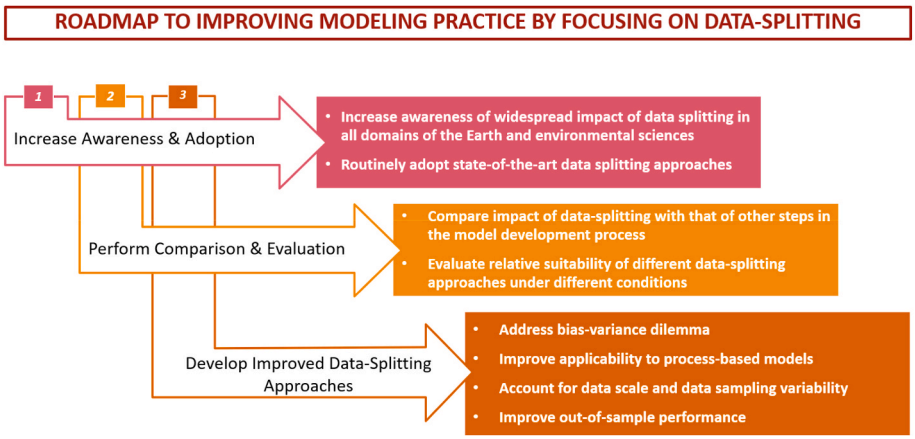


Fig. 3. Roadmap to improve data-splitting as part of the model development and evaluation process – the elephant in the room.

data in the final calibration, provided the first step indicates that overfitting is not an issue (Mai, 2023).

b. **Improving the applicability of data-splitting methods to process-based models.** As discussed, many existing data-splitting methods do not address the need to maintain the underlying time structure of the data, and hence cannot be applied directly to process-based models that require time-consecutive observations as inputs. Consequently, efforts need to be intensified to develop strategies that extend the application of data-splitting methods to process-based model development (see Zheng et al., 2022 for an example).

c. **Accounting for data scale and data sampling variability.** When considering data sampling variability, it is important to bear in mind the scale of data (i.e., big data vs small data) and the complexity of the system being modeled. Small data represent a particular challenge for data splitting due to the limited scale of data and further efforts are required to provide guidance on which data-splitting approach is most appropriate. The use of cross-validation is likely to be advantageous in these settings, as multiple splits are made, exhausting all combinations of model development and validation splits, although this introduces new potential concerns. For instance, by having multiple calibration data subsets, different model parameters will be produced for each split. However, no matter how extensive the available observations, the available data necessarily represent a specific realization of the underlying physical process. This means that data sampling variability (i.e., variability due to which particular subset of the available data is used for model development and evaluation (see Zheng et al. (2023)) inevitably exists, even in large datasets, and again this issue has been largely ignored by the modeling community. Consequently, when developing models, it is necessary to properly account for data sampling variability, and this should be a focus of future research.

d. **Improving out-of-sample-performance.** Whereas this position paper focuses on new methods to improve "in-sample" model performance, future work should also address the need to better use the available data to enhance out-of-sample model performance. In principle, evaluating the performance of models in "true" out-of-sample prediction is inherently challenging, as it involves extrapolating system behaviors into unobserved state spaces. For instance, models are often employed to address questions such as the potential impact of changes in climate on system performance (e.g., Wu et al., 2023). To advance this field, future research should concentrate on developing novel methods within the following two interrelated approaches:

   i. **Stress-testing models.** One area of focus for future research involves stress-testing models to evaluate their performance when subjected to inputs beyond the range of data used for standard calibration and validation. This approach involves creating hypothetical input data by perturbing specific statistical properties of these data to stress test the model (e.g., Bennett et al., 2021; Culley et al., 2021; Razavi, 2021). For instance, historical temperature time series in a region can be shifted to simulate warmer climates, which can then be utilized as input for a hydrologic model. By observing the model's response to such stressors, valuable insights can be gained into its performance in true out-of-sample prediction scenarios.

   ii. **Model intercomparison in extrapolation scenarios.** In cases where multiple alternative models exist for a system, comparing their responses under stress-testing conditions and benchmarking those responses against the available knowledge base can provide valuable insights into model behaviors under unseen conditions (e.g., see out-of-sample validation of 12 process-based and 1 data-driven model in Mai et al. (2022)). This allows for a comprehensive evaluation of how different models perform in extrapolation scenarios. However, it is crucial to note that generating hypothetical inputs by perturbing historical data may pose challenges, especially for process-based models. These models often require the preservation of certain relationships between different input variables (e.g., between precipitation and air humidity) during the data generation and stress-testing processes. This requirement demands careful attention and consideration to ensure the integrity of the data used in stress testing these models (see Guo et al. (2018) and Culley et al. (2019)).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (52261160379). Wenyan Wu acknowledges support from the Australian Research Council via the Discovery Early Career Researcher Award (DE210100117).

References

Addor, N., Melsen, L.A., 2019. Legacy, rather than adequacy, drives the selection of hydrological models. Water Resour. Res. 55, 378–390.

Arsenault, R., Brissette, F., Martel, J.L., 2018. The hazards of split-sample validation in hydrological model calibration. J. Hydrol. 566, 346–362.

Ascough II, J.C., Maier, H.R., Ravalico, J.K., Strudley, M.W., 2008. Future research challenges for incorporation of uncertainty in environmental and ecological decision-making. Ecol. Model. 219 (3–4), 383–399.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S. A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andreassian, V., 2013. Characterising the performance of environmental models. Environ. Model. Software 40, 1–20.

Bennett, B., Devanand, A., Culley, S., Westra, S., Guo, D., Maier, H.R., 2021. A modelling framework and R package for evaluating system performance under hydroclimate variability and change. Environ. Model. Software 139, 104999.

Bowden, G.J., Maier, H.R., Dandy, G.C., 2002. Optimal division of data for neural network models in water resources applications. Water Resour. Res. 38 (2), 1010.

Chen, J., Zheng, F., May, R., Guo, D., Gupta, H., Maier, H.R., 2022. Improved data splitting methods for data-driven hydrological model development based on a large number of catchment samples. J. Hydrol. 613 (Part A), 128340.

Coron, L., Andreassian, 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 Resour. Res. 48 (5), W05552.

Craven, P., Wahba, G., 1978. Smoothing noisy data with spline functions. Numer. Math. 31, 377–403.

Culley, S., Bennett, B., Westra, S., Maier, H.R., 2019. Generating realistic perturbed hydro-meteorological time series to inform scenario-neutral climate impact assessments. J. Hydrol. 576, 111–122.

Culley, S.A., Maier, H.R., Westra, S., Bennett, B., 2021. Identifying critical climate conditions for use in scenario-neutral climate impact assessments. Environ. Model. Software 136, 104948.

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

Gauch, M., Mai, J., Lin, J., 2021. The proper care and feeding of CAMELS: how limited training data affects streamflow prediction. Environ. Model. Software 135, 104926.

Gibbs, M., McInerney, D., Humphrey, G., Thyer, M.A., Maier, H.R., Dandy, G.C., Kavetski, D., 2018. State updating and calibration period selection to improve dynamic monthly streamflow forecasts for an environmental flow management application. Hydrol. Earth Syst. Sci. 22 (1), 871–887.

Guo, D., Westra, S., Maier, H.R., 2018. An inverse approach to perturb historical rainfall data for scenario-neutral climate impact studies. J. Hydrol. 556, 887–890.

Guo, D., Zheng, F., Gupta, H., Maier, H.R., 2020. On the robustness of conceptual rainfall-runoff models to calibration and evaluation data set splits selection: a large sample investigation. Water Resour. Res. 56 (3), e2019WR026752.

Gupta, H.V., Sorooshian, S., Yapo, P.O., 1998. Towards improved calibration of hydrologic models: multiple and non-commensurable measures of information. Water Resour. Res. 34 (4), 751–763.

Gupta, H.V., Perrin, C., Kumar, R., Blöschl, G., Montanari, A., Kumar, R., Clark, M., Andreassian, V., 2014. Large-sample hydrology: a need to balance depth with breadth. Hydrol. Earth Syst. Sci. 18, 1–15.

Hamilton, S., Pollino, C.A., Stratford, D.S., Fu, B., Jakeman, A.J., 2022. Fit-for-purpose environmental modeling: targeting the intersection of usability, reliability and feasibility. Environ. Model. Software 148, 105278.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, second ed. Springer New York. https://doi.org/10.1007/978-0-387-84858-7.

Hoell, A., Funk, C., Barlow, M., Cannon, F., 2017. A physical model for extreme drought over Southwest Asia. Climate Extremes Patterns and Mechanisms 226, 283–298.

Hrachowitz, M., Savenije, H.H.G., Blöschl, G., McDonnell, J.J., Sivapalan, M., Pomeroy, J.W., Arheimer, B., Blume, T., Clark, M.P., Ehret, U., Fenicia, F., Freer, J. E., Gelfan, A., Gupta, H.V., Hughes, D.A., Hut, R.W., Montanari, A., Pande, S., Tetzlaff, D., Troch, P.A., Uhlenbrook, S., Wagener, T., Winsemius, H.C., Woods, R.A., Zehe, E., Cudennec, C., 2013. A decade of predictions in ungauged basins (PUB)—a review. Hydrol. Sci. J. 58 (6), 1198–1255.

Humphrey, G.B., Maier, H.R., Wu, W., Mount, N.J., Dandy, G.C., Abrahart, R.J., Dawson, C.W., 2017. Improved validation framework and R-package for artificial neural network models. Environ. Model. Software 92, 82–106.

Jakeman, A.J., Letcher, R.A., Norton, J.P., 2006. Ten iterative steps in development and evaluation of environmental models. Environ. Model. Software 21, 602–614.

Klemes, V., 1986. Operational testing of hydrological simulation models. Hydrol. Sci. J. 31 (1), 13–24.

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrol. Earth Syst. Sci. 22, 6005–6022.

Lees, T., Reece, S., Kratzert, F., Klotz, D., Gauch, M., De Bruijn, J., Sahu, R.K., Greve, P., Slater, L., Dadson, S.J., 2022. Hydrological concept formation inside long short-term memory (LSTM) networks. Hydrol. Earth Syst. Sci. 26, 3079–3101.

Lever, J., Krzywinski, M., Altman, N., 2016. Points of significance: model selection and overfitting. Nat. Methods 13 (9), 703–705.

Li, C., Zhang, L., Wang, H., Zhang, Y.Q., Yu, F.L., Yan, D.H., 2012. The transferability of hydrological models under nonstationary climatic conditions. Hydrol. Earth Syst. Sci. 16 (4), 1239–1254.

Liu, D., Guo, S., Wang, Z., Liu, P., Yu, X., Zhao, Q., Zou, H., 2018. Statistics for sample splitting for the calibration and validation of hydrological models. Stoch. Environ. Res. Risk Assess. 32, 3099–3116.

Mai, J., 2023. Ten strategies towards successful calibration of environmental models. J. Hydrol. 620 (A), 129414.

Mai, J., Shen, H., Tolson, B.A., Gaborit, É., Arsenault, R., Craig, J.R., Fortin, V., Fry, L.M., Mgauch, M., Klotz, D., Kratzert, F., O'Brien, N., Princz, D.G., Rasiya Koya, S., Roy, T., Seglenieks, F., Shrestha, N.K., Temgoua, A.G.T., Vionnet, V., Waddell, J.E., 2022. The great lakes runoff intercomparison project phase 4: the great lakes (GRIP-GL). Hydrol. Earth Syst. Sci. 26 (13), 3537–3572.

Maier, H.R., Jain, A., Dandy, G.C., Sudheer, K.P., 2010. Methods used for the development of neural networks for the prediction of water resource variables in river systems: current status and future directions. Environ. Model. Software 25 (8), 891–909.

Maier, H.R., Galelli, S., Razavi, S., Castelletti, A., Rizzoli, A., Athanasiadis, I.N., Sanchez-Marre, M., Acutis, M., Wu, W., Humphrey, G.B., 2023. Exploding the myths: an introduction to artificial neural networks for prediction and forecasting. Environ. Model. Software 167, 105776.

May, R.J., Maier, H.R., Dandy, G.C., 2010. Data splitting for artificial neural networks using SOM-based stratified sampling. Neural Network. 23 (2), 283–294.

Mount, N.J., Maier, H.R., Toth, E., Elshorbagy, A., Solomatine, D., Chang, F.-J., Abrahart, R.J., 2016. Data-driven modelling approaches for social-hydrology: opportunities and challenges within the panta rhei science plan. Hydrol. Sci. J. 61 (7), 1192–1208.

Pianosi, F., Beven, K., Freer, J., Hall, J.W., Rougier, J., Stephenson, D.B., Wagener, T., 2016. Sensitivity analysis of environmental models: a systematic review with practical workflow. Environ. Model. Software 79, 214–232.

Picard, R.R., Berk, K.N., 1990. Data splitting. Am. Statistician 44 (2), 140–147.

Raupach, M., Briggs, P., Haverd, V., King, E.A., Paget, M.J., Trudinger, C.M., 2009. Australian Water Availability Project (AWAP): CSIRO Marine and Atmospheric Research Component: Final Report for Phase 3. Aspendale. CSIRO Marine and Atmospheric Research procite:4175c8ca-86e1-496a-96c5-13dc53e2cf1d. http://hdl.handle.net/102.100.100/115368?index=1.

Raupach, M., Briggs, P., Haverd, V., King, E., Paget, M., Trudinger, C., 2012. Australian Water Availability Project. CSIRO Marine and Atmospheric Research, Canberra, Australia. Retrieved from. http://www.csiro.au/awap.

Razavi, S., 2021. Deep learning, explained: fundamentals, explainability, and bridgeability to process-based modelling. Environ. Model. Software 144, 105159.

Razavi, S., Jakeman, A., Saltelli, A., Prieur, C., Iooss, B., Borgonovo, E., Plischke, E., Lo Piano, S., Iwanaga, T., Becker, W., Tarantola, S., Guillaume, J.H.A., Jakeman, J., Gupta, H., Melillo, N., Rabitti, G., Chabridon, V., Duan, Q., Sun, X., Smith, S., Sheikholeslami, R., Hosseini, N., Asadzadeh, M., Puy, A., Kucherenko, S., Maier, H. R., 2021. The future of sensitivity analysis: an essential discipline for systems modeling and policy support. Environ. Model. Software 137, 104954.

Saltelli, A., Jakeman, A., Razavi, S., Wu, Q., 2021. Sensitivity analysis: a discipline coming of age. Environ. Model. Software 146, 105226.

Seibert, J., McDonnell, J.J., 2015. Gauging the ungauged basin: relative value of soft and hard data. J. Hydrol. Eng. 20 (1), A4014004.

Shen, H., Tolson, B.A., Mai, J., 2022. Time to update the split-sample approach in hydrological model calibration. Water Resour. Res. 58, e2021WR031523.

Singh, S.K., Bárdossy, A., 2012. Calibration of hydrological models on hydrologically unusual events. Adv. Water Resour. 38, 81–91.

Snee, R.D., 1977. Validation of regression models: methods and examples. Technometrics 19 (4), 415–428.

Trotter, L., Saft, M., Peel, M.C., Fowler, K.J.A., 2023. Symptoms of performance degradation during multi-annual drought: a large-sample, multi-model study. Water Resour. Res. 59, e2021WR031845.

Vaze, J., Post, D.A., Chiew, F.H.S., Perraud, J.-M., Viney, N.R., Teng, J., 2010. Climate non-stationarity – validity of calibrated rainfall–runoff models for use in climate change studies. J. Hydrol. 394 (3–4), 447–457.

Vilas, M.P., Egger, F., Adams, M.P., Maier, H.R., Robson, B., Mestres, J.F., Stewart, L., Maxwell, P., O'Brien, K.R., 2023. TALKS: a systematic framework for resolving model-data discrepancies. Environ. Model. Software 163, 105668.

Van Vliet, J., Bregt, A.K., Brown, D.G., van Delden, H., Heckbert, S., Verburg, P.H., 2016. A review of current calibration and validation practices in land-change modelling. Environ. Model. Software 82, 174–182.

Vrugt, J.A., Bouten, W., Gupta, H.V., Sorooshian, S., 2003. Correction to "Toward improved identifiability of hydrologic model parameters: the information content of experimental data". Water Resour. Res. 39 (3), 10–11.

Wöhling, T., Schöniger, A., Gayler, S., Nowak, W., 2015. Bayesian model averaging to explore the worth of data for soil-plant model selection and prediction. Water Resour. Res. 51, 2825–2846.

Wu, W., May, R.J., Maier, H.R., Dandy, G.C., 2013. A benchmarking approach for comparing data splitting methods for modeling water resources parameters using artificial neural networks. Water Resour. Res. 49 (11), 7598–7614.

Wu, W., Dandy, G.C., Maier, H.R., 2014. Protocol for developing ANN models and its application to the assessment of the quality of the ANN model development process in drinking water quality modelling. Environ. Model. Software 54, 108–127.

Wu, W., Eamen, L., Dandy, G., Razavi, S., Kuczera, G., Maier, H.R., 2023. Beyond engineering: a review of reservoir management through the lens of wickedness, competing objectives and uncertainty. Environ. Model. Software 167, 105777.

Yang, J., Jakeman, A., Fang, G., Chen, X., 2018. Uncertainty analysis of a semi-distributed hydrologic model based on a Gaussian process emulator. Environ. Model. Software 101, 289–300.

Zheng, F., Maier, H.R., Wu, W., Dandy, G.C., Gupta, H.V., Zhang, T., 2018. On lack of robustness in hydrological model development due to absence of guidelines for

selecting calibration and evaluation data: demonstration for data-driven models. Water Resour. Res. 54 (2), 1013–1030.
Zheng, F., Chen, J., Maier, H.R., Gupta, H., 2022. Achieving robust and transferable performance for conservation-based models of dynamical physical systems. Water Resour. Res. 58 (5), 17406751.
Zheng, F., Chen, J., Ma, Y., Chen, Q., Maier, H.R., Gupta, H.V., 2023. A robust strategy to account for data sampling variability in the development of hydrological models. Water Resour. Res. 59 (3), e2022WR033703.