Responses to short comments by referee #2

Dear Editor, dear Referee,

We thank the second referee for the review of our manuscript and the detailed comments. We will in the following reply to the comments point by point. The Referee comments are in blue.

Comment 1: Summary: As the title suggests, the authors present a practical approach to evaluate model computational complexity vs. model performance.

Comment 2: Comments: this is suitable technical note. I however would not so easily agree with the authors’ claim that their approach is a general approach if they mean in terms of its theoretical underpinning. General perhaps in terms of how easily it can be applied. Nonetheless this is an interesting article that readers can learn from and apply the methodology in diverse settings. Following are my comments in detail.

Reply 2: Thank you for this comment, which made us aware that we have to better explain the meaning of 'general' in our manuscript. It is used in two different but related contexts: The first refers to the use of 'strace' as a tool to measure computational complexity of a model. It is general in the sense that i) within its domain of application (computer-based models), it can be applied to all models, whether or not the source code is available, and ii) it is sensitive to all computational aspects of a model (the amount of required forcing data, the size of the model itself, its time stepping, spatial resolution, numerical scheme, etc.).

The second use refers to the 'bit by bit' concept: It is general within its domain of application (evaluation of models in the process of science) in the sense that i) it covers all key aspects (performance and complexity), and ii) that by expressing all aspects in unit [bit], it allows their combination and comparison. This would not be possible if e.g. performance were expressed as RMSE, and computational complexity in [s].

In a revised version of the manuscript, we will better explain the meaning of 'general'.

Comment 3: How max parsimony + max performance -> max generalizability? Theoretical rigor behind the claim is missing

Reply 3: This claim refers to the idea that when describing data well with a simple model, it tends to work well for prediction of unseen data. This is building on Occam's razor. The theoretical basis of this claim is addressed in detail in Weijs and Ruddell (2020) and references therein. This is widely accepted in practice and, for example, visible in the Akaike Information Criterion (a performance term + a complexity penalization). Algorithmic information theory provides a deeper theoretical basis, and explains that shorter descriptions are more likely to be generalizable (predictive). The theoretical rigor draws upon many concepts beyond the typical expertise of the intended audience (universality of computation, self-delimiting Turing machines etc.), that we do not think are helpful to discuss here as it is not the main focus of the paper. Hence we refer to the WRR paper that in turn refers to the AIT literature, such as the convergence results by Solomonoff (1978).

On an intuitive level, the idea is that we can see physical processes as performing computations, simple computations are more likely to arise naturally, and thus have a higher a priori probability (see algorithmic probability). This is a quantification of Occam's razor and expresses a belief of structure in the universe. Occam’s razor is an essential part of finding descriptions that generalize beyond the given data set, as without it there are infinitely many equally valid explanations. The explanations that are simple tend to apply more easily to new situations and thus are more generalizable. This is the reason why inductive inference works at all, and why intelligent beings tend to have an evolutionary advantage in this universe.

One more note about theoretical rigor: Even with AIT, we cannot prove that inductive inference works, but we can better formally describe and measure how well it works in practice (so the only evidence that induction works is itself an inductive inference).

Comment 4: Concept of information loss assumes full specification of the data generating process, which often is not the case in hydrological modeling. Please elaborate further how this is dealt with

Reply 4: Thank you for this comment, which shows that we need to better explain the usage of 'information loss'. We have two replies:

- The referee correctly states that the process generating the observations we have is typically not fully known in earth science problems, except for virtual reality settings. This is the case no matter whether we measure performance by RMSE, NSE, or information loss. So the best we can do is to measure performance against the observations we have, and hope that it allows conclusions that are also valid for the true underlying data generating system. In that sense, we can use the distribution and dependencies of the data as a benchmark, against which information losses can be calculated.
- Alternatively, if we want to emphasize the limited information contained in the data about the system (e.g. in cases of very few observations), we can start from the opposite end: We can specify maximum entropy (=minimum prior knowledge) distributions for all system variables of interest, and then measure the information gain of the available data against this flat prior. In such a case, the y-axis in figure 3 shows information gain, but the overall interpretation remains the same.

Weijs and Ruddell (2020), which we refer to throughout the text, use information losses because they directly translate to a description length. For reasons of comparability we prefer to stick to the same measure in the manuscript. This has the advantage that all measures are negatively oriented scores, where lower is better. We will add a related explanation to a revised version of the manuscript.

Comment 5: In this context thinking of models as compression algorithms of data is shallow in its treatment of what complexity means in terms of learning from patterns, especially when patterns are generated from complex data generating process. I can understand the concept being a good one in describing model complexity appropriately when we know the data generating process and are playing with its approximations and trading that off with information loss incurred by the approximations. So the authors claim of universality is overdone in real world hydrological systems, perhaps it may work in Shannon's communication systems.

Reply 5: We would like to reply to this comment in two parts: We think algorithmic information theory is far from shallow, but we do agree there are still challenges to be overcome for its practical application in hydrology.

Since descriptive complexity is not the main focus of this paper, we did not go into much detail, but more discussion on this can be found in the WRR paper on Occam's Razor. One thing to note is that however complex a hydrological system may be, all its processes can – in principle – be simulated by a Universal Turing Machine. There are connections at the deepest level between physics, information, and computation (See e.g. the Church-Turing-Deutsch thesis, black hole thermodynamics, the Landauer's principle), therefore AIT certainly has bearing on inference outside the world of computer science and communication systems. Whatever real world data set is produced by whatever generating process, it will have a finite Kolmogorov complexity, and in principle a binary computer program could perfectly reproduce it. Approaching that Kolmogorov complexity with finding shortest descriptions of all dependencies in the data is exactly the same process as modeling the patterns in that data. Note that there is no reference to an underlying data generating mechanism here, only to the data itself, which in the end is our only window on reality. Now of course our hope is that the model, or compression algorithm, found has some functional similarity with the data generating process, and we can use it for predictions. And that usually works, because the universe typically is ruled by a decent bit of order and simplicity (with the exception of my desk).

Now as for practical application, even before theoretical computability issues come in, practical limitations will make it impossible to pursue finding the full likeliest algorithm that produced the data. Especially since the typical data sets used in hydrology do not fully specify the system, or even do not specify everything we know about it. However, there are several ways in which this lack of completeness can be dealt with by describing the variations unexplained by the patterns we found literally (which is the equivalent of viewing the data generating process as stochastic).

Comment 6: Related to the above, it is for this reason that synthetic cases may be easier to demonstrate. Author’s claim to universality should first provide a rigorous theoretical treatment that has not even been provided in the WRR paper that the authors allude to.

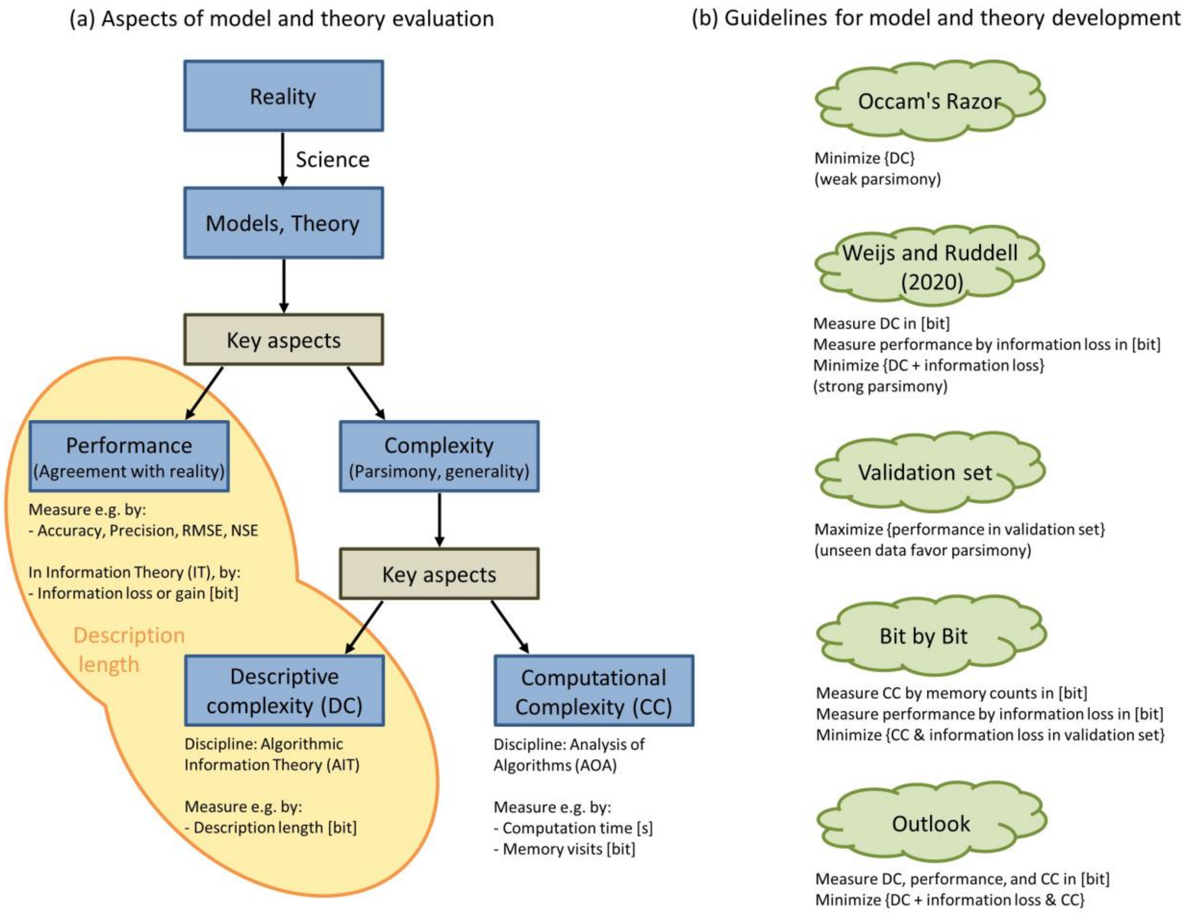
Reply 6: Respectfully, we do not understand the referee's point here: In the manuscript we provide applications of typical hydrological models to real-world data. However, we understood that we should use the term 'universality' more carefully. In a revised version of the manuscript, we will replace 'universality' with 'generality', and better explain the use of the latter (see our reply to comment 2).

Comment 7: That is the reason why the authors attempt to extend it to real world case studies is not constructive unless the error model of the residuals is completely specified (or known).

Reply 7: We are not sure we understand the referee's concern here. We assume that the mentioned error model of the residuals specifies the disagreement between the data generating system and the observations thereof. If this is the case, please see our reply to comment 4. If this is not the case, please explain.

Comment 8: I am not at all clear how computational complexity is linked to inference. This is where the paper lost me in its attempt to connect this paper to their earlier WRR paper. Here while authors talk about inference without reference to predictive performance, no clear theory on how computational complexity is linked to generalizability is given.

Reply 8: Thank you for this comment. It shows that we need to better explain the relation among the concepts introduced or discussed in the manuscript. In order to increase overall clarity, we suggest adding a figure and related discussion to the manuscript (see below).


(a) Aspects of model and theory evaluation
Reality
Science
Models, Theory
Key aspects
Performance
(Agreement with reality)
Measure e.g. by:
- Accuracy, Precision, RMSE, NSE
In Information Theory (IT), by:
- Information loss or gain [bit]
Description length
Complexity
(Parsimony, generality)
Key aspects
Descriptive complexity (DC)
Discipline: Algorithmic Information Theory (AIT)
Measure e.g. by:
- Description length [bit]
Computational Complexity (CC)
Discipline: Analysis of Algorithms (AOA)
Measure e.g. by:
- Computation time [s]
- Memory visits [bit]
(b) Guidelines for model and theory development
Occam's Razor
Minimize {DC}
(weak parsimony)
Weijs and Ruddell (2020)
Measure DC in [bit]
Measure performance by information loss in [bit]
Minimize {DC + information loss}
(strong parsimony)
Validation set
Maximize {performance in validation set}
(unseen data favor parsimony)
Bit by Bit
Measure CC by memory counts in [bit]
Measure performance by information loss in [bit]
Minimize {CC & information loss in validation set}
Outlook
Measure DC, performance, and CC in [bit]
Minimize {DC + information loss & CC}

In short, we will discuss:

- Weijs and Ruddell (2020) extend Occam's Razor – which evaluates descriptive complexity only – by additionally measuring performance. Also, they measure both criteria in unit [bit], which allows a joint treatment and comparison.
- Classical validation set approaches seek to maximize model performance on unseen data. By using unseen data, general/parsimonious model approaches are favored, and overfitting is avoided. The approach by Weijs and Ruddell (2020) assures generality/parsimony by including descriptive complexity into their maximization function. For the latter we will include references to relevant literature from Algorithmic Information Theory.
- The bit by bit approach adds another key dimension of model complexity: computational complexity, measured by 'strace' in unit [bit]. It is important to note that computational complexity is neither a surrogate for measuring performance, nor for measuring descriptive complexity. In order to provide a comprehensive evaluation of a model, computational complexity can be either combined with the approach by Weijs and Ruddell (2020), OR with a validation set approach. In both cases, there is a guard for overfitting (descriptive complexity in the first, evaluation on unseen data in the second). In the current version of the manuscript, we combined computational complexity with performance on a data set seen

through calibration - as in Weijs and Ruddell (2020) - but did not add a descriptive complexity control. This is inconsistent with the above explanation. We therefore suggest that in a revised version of the paper, we will use computational complexity in combination with performance measured by information loss in a validation set approach. We hope that readers from the hydrological community will find this easy to relate to as it follows established procedures of cal/val.

- We will further describe an approach - and suggest it for future research - where computational complexity is combined with the approach by Weijs and Ruddell (2020). This avoids the need for splitting available data into cal and val subsets, but still favors general/parsimonious models. In addition, as all criteria can be measured in unit [bit], this allows a joint treatment and comparison of all key aspects of model evaluation.

We hope that by the figure and the explanation, the relation between the different guiding principles, and their relation to the scientific process of model/theory building/evaluation becomes more clear. We suggest keeping the existing in-depth introduction to the theory, as placing the bit by bit approach into this framework is one main goal of the paper.

A direct reply to the referee comment: Computational complexity is not directly linked to inference, rather it provides another key facet of model evaluation that can be used in addition to the criteria used by Weijs and Ruddell (2020). Also, we do not claim that computational complexity provides guidance on the generality of models or laws, our claim is rather that measuring computational complexity by 'strace' and the 'bit by bit' approach are generally applicable (please see also our replies to comment 2).

Comment 9: Even if ’generalization’ laws have been found, how good they are depend on how well they hold on unseen data, i.e. predictive uncertainty

Reply 9: Please see our reply to comment 8 about how the different concepts discussed in the manuscript (Occam's razor, Weijs and Ruddell 2020, Validation set, bit by bit) assure that general models are favored.

Comment 10: I was totally lost in the philosophical arguments at the end of the introduction paper. Please delete, it appears to have been placed to impress the reader. I am reacting to it in quite an opposite manner

Reply 10: We agree that the discussion about the fundamental natural complexity (lines 98-112 in the manuscript) is not strongly connected to the rest of the manuscript. Nevertheless, we think it is important to mention that all the concepts discussed throughout the manuscript measure only facets of the complexity of the natural system under investigation. In a revised version of the manuscript, we will shorten this section.

Comment 11: The way Prob for entropy measure has been calculated is in itself a model that depends on the choice and number of bins. That has implications for how well Prob has been estimated from limited data in terms of how such frequency estimates converge to true Prob (ie it has its own complexity challenges) that the approach so very much relies on. Perhaps this can be discussed in bit more detail.

Reply 11: Agreed. We will add a short discussion about the influence of binning choices, and point to related literature.

Finally two major comments:

Comment 12: the authors should show predictive performance to demonstrate generalizability. Or validation, even if in narrative form, by comparing their conclusions with what other authors, not linked to information theory applied to water, have said.

Reply 12: Please see our replies to comment 8.

Comment 13: the authors again need to place their finding in the landscape of other complexity studies, especially in modeling MOPEX catchments, in hydrology. How do their conclusions regarding complexity compare with the narrative presented here? This will only add value to an already large literature set of hydrological model complexity, esp wrt to streamflow.

Reply 13: Agreed. The usages of the term 'complexity' are manifold – not only but also in hydrology - and to date no unique definition exists. For clarification, and to put our work into context, we will add to the manuscript a brief overview on its usage in hydrology, and refer more extensively to prior work on model complexity in hydrology.

Yours sincerely,

Uwe Ehret, on behalf of all co-authors