Hydrol. Earth Syst. Sci. Discuss.,
https://doi.org/10.5194/hess-2020-487-AC2, 2020

Hydrology and Earth System Sciences
Open Access
Discussions

Interactive comment on “Hydrologically Informed Machine Learning for Rainfall-Runoff Modelling: Towards Distributed Modelling” *by* Herath Mudiyanselage Viraj Vidura Herath et al.

Herath Mudiyanselage Viraj Vidura Herath et al.

e0149661@u.nus.edu

Received and published: 24 December 2020

We would like to express our gratitude towards the time and effort that the editor and all the reviewers dedicated to providing valuable feedback to help in improving this journal paper. We appreciate the insightful comments and suggestions given by the reviewers on this paper.

We are pleased that both the reviewers have accepted our proposed toolkit as a novel, valuable contribution to the field of hydrological modelling. As most of the concerns raised by the reviewers are related to the structure of the manuscript rather than the content, we believe a restructured manuscript would adequately address all comments

and concerns of reviewers.

Overall, we have addressed all the concerns of the reviewers through this document. Here, we have included the point-by-point response to the reviewers’ comments and concerns. The original reviewer’s comment is marked by starting the line with “>”, while the corresponding response is annotated with “[Response]”. The line numbers in reviewers’ comments refer to the original manuscript.

Reviewer 2:

GENERAL COMMENTS

> labelled as MIKA-SHA, though I would assume it should be MIK-ASHA based on the full title of the framework on Line 18 “Machine Induction Knowledge-Augmented System Hydrologique Asiatique”

[Response] We have placed the “-” in a wrong position. We want to label our toolkit as MIKA-SHA. Hence the name should be “Machine Induction Knowledge Augmented – System Hydrologique Asiatique”. We will correct this in the revised manuscript.

> These first pages provide a very basic review of well-established topics in hydrological modelling (e.g. physics based vs data based modes, fixed v flexible, lumped v distributed, etc.). I think the authors’ intentions are to provide some baseline definitions and maybe highlight some common issues under each subtheme. However, in its current form, it reads like a very “high level” summary (similar to textbook descriptions) rather than adding value or providing the current state of research in each field. The way this initial text is framed suggests that the proposed framework will collectively address many of the pitfalls found in existing data-driven methods, whereas in fact the proposed method has a very specific focus (more details on this are below). Many of the problems with hydrological models (physics or data-driven) are well-documented and it would be better if the authors focus on some recent innovations in these fields, and used these advances as a basis of comparison for their proposed method (some

suggestions below). This would enable a proper or more detailed evaluation of the proposed method, which currently seems lacking. I would recommend reducing the focus on the introductory text and expanding the methods sections (Section 4 and 5) to better explain the overall method (much of it is in bullet form). Lastly, the analysis is performed on a single watershed, but the discussion of the results seems to imply that the advantages of the proposed methods are more widespread (which may indeed be the case), but I think a comparison with other watersheds may be beneficial for to support these statements.

[Response] As you have correctly stated here, our intension was to provide basic background on different hydrological modelling strategies as our toolkit is a hybrid approach which is not founded on a single modelling strategy. We use GP as a data-driven technique while flexible modelling frameworks and semi-distributed modelling paradigm are used as theory-driven techniques. We did not provide in-depth discussions on the above topics in our manuscript. Instead, for more details, we navigated interested readers to relevant papers and textbooks through citations.

As you have suggested, we will expand the methods section by providing more details about our proposed toolkit in the revised manuscript. Further, we will revise the introductory section while focusing more on recent developments in each field. The focus of the MIKA-SHA will be explicitly stated in the revised manuscript.

Indeed, we have tested MIKA-SHA on many other catchments and it provides equally good results with them. However, in this manuscript, our main objective was to introduce the toolkit rather than focusing more on its applications. We limited the result section into one catchment just to show how MIKA-SHA operates. The unique feature of our toolkit is the readily interpretable nature of induced models. Hence, we allocated some space to demonstrate how induced models can be explained along with catchment characteristics and previous research findings. Therefore, if we to add more catchments, the length of the paper would be increased significantly. However, in our future work, we will focus on applications of MIKA-SHA.

SPECIFIC COMMENTS

Abstract

> Machine learning and data sciences have been extensively researched in hydrology – I do not agree that "in general,: : :limited use in scientific fields". I think a better description would be that perhaps they are not put in practice as much as they are used for research applications. However, support for such a statement would be difficult to quantify. Additionally, "commercial fields" is a vague term that should be better described.

[Response] Yes, we agree that machine learning and data science have been extensively researched in hydrology. However, we made the above statement based on the following arguments.

"Unfortunately, this notion of black-box application of data science has met with limited success in scientific domains" – Karpatne et al. (2017).

"There are two primary characteristics of knowledge discovery in scientific disciplines that have prevented data science models from reaching the level of success achieved in commercial domains." – Karpatne et al. (2017).

"We suggest that there is a potential danger to the hydrological sciences community in not recognizing how transformative machine learning will be for the future of hydrological modeling." – Nearing et al. (2020).

"As mentioned above, the potential advantage with ML (over process-based models), is that we might be able to train DL models to react to exogenous inputs that are only indirectly related to the states and fluxes of a hydrology model (e.g., directly ingest energy or commodity prices as proxies for water demand). We see significant potential here, but as far as we know this has not been tested using state-of-the-art DL." – Nearing et al. (2020).

"Compared to some other disciplines, hydrology has not witnessed the wide use of DL.

Applications of DL have been few, especially in a big data setting, and the list of papers reviewed in this subsection is exhaustive to the author's best compilation effort." - Shen (2018).

> FUSE and SUPERFLEX are mentioned in the abstract without any contextual information: what are these frameworks?

[Response] These are the two flexible rainfall-runoff modelling frameworks that we use as the existing hydrological knowledge to guide the GP algorithm. As it would be confusing to the readers, the two frameworks will be introduced in the introduction section and will be removed from the abstract in the revised manuscript.

> Very limited information, other than a general description of the framework and its advantages, are presented in the abstract. It would be useful for a reader to better understand how the model works, how well the model works, and the basis for the conclusion that the model is ": : :compatible with previous findings".

[Response] We will add a few sentences to the abstract about how the toolkit works, how well it performs, and the basis for the above conclusion in the revised manuscript.

Introduction

> "Hydrological models play a key role in capturing the discharge signatures of watersheds." I am not sure if this statement is really necessary – in some ways it's obvious, and in others, I am not sure if it is correct: models are used to simulate the discharge in a watershed; underlying data provides the mechanism to understand the system. By discharge signatures, do the authors mean the factors driving the discharge?

[Response] He used the term signature to refer to catchment runoff dynamics. MIKA-SHA toolkit can also be used to identify the discharge driving factors by using it separately with HRUs defined based on topography, soil type and lithology of the catchment of interest. What we meant by the above sentence is hydrological models are important in understanding the runoff dynamics of a catchment. Once a model is built/ calibrated,

it can be used for quantitative extrapolation or prediction that will hopefully be helpful in decision making.

> "So far, no hydrological model can perform equally well over the entire range of problems." Unclear what the authors mean by problems here. But if the message is that no hydrological model can simulate a system under the full range of observed conditions, perhaps some supporting information is needed.

[Response] The idea that we want to stress here can be summarized as follows.

Each hydrological model is built on a perceptual model which originates in the mind of the model developer. As stated by Beven (2012), "The perceptual model is the summary of our perceptions of how the catchment responds to rainfall under different conditions or, rather, your perceptions of that response. A perceptual model is necessarily personal."

However, evidence suggests that catchments tend to behave uniquely due to their spatial heterogeneity.

Therefore it is not reasonable to assume a model built on one hypothesis to perform equally well in different conditions where the underlying physical phenomenon may be different from the assumed hypothesis.

Our statement was also based on the following literature.

" Whenever detailed studies of flow pathways are carried out in the field we find great complexity. This complexity is one reason why there is no commonly agreed modelling strategy for the rainfall–runoff process but a variety of options and approaches that will be discussed in the chapters that follow." – Beven (2012).

"Given these considerations, at least in the context of rainfall-runoff modeling, we do not see a strong a priori reason why the dynamics of catchments that are widely different in their geomorphology and climatology should always reduce to a single common form at coarsely lumped scales." – Fenicia et al. (2011).

> “This leads to different research directions seeking different hydrological models based on different modelling strategies.” Again, this is a pretty vague statement and not useful in outlining the present research. Are the authors trying to state that several hydrological modelling approaches are currently being researched to solve different research problems?

[Response] As mentioned above, our toolkit is a novel hybrid modelling approach founded on both data-driven and theory-driven techniques. One may argue that is there a necessity for developing new modelling approaches as we already have an overwhelming number of hydrological models. Therefore, we used the above sentence to highlight that there is still a necessity to develop novel and innovative modelling approaches in hydrological modelling.

“there is no commonly agreed modelling strategy for the rainfall–runoff process but a variety of options and approaches” – Beven (2012).

“It seems unlikely that all these different reasons for using models can be met by a single modelling approach in the future. The next generation of hydrological models is likely therefore to continue to be diverse.” – Beven (2012).

> “Therefore, the final goal of any successful hydrological model must be based on a physically meaningful model architecture along with a good predictive performance.” This also seems fairly obvious and standard practice in model development and selection.

[Response] Yes, this would be the standard practice in model development and selection. However, a common criticism of machine learning (data-driven) is that it lacks understanding or explainability of their model architectures. On the other hand, physics-based models which have physically meaningful architectures based on the understanding of the runoff processes often fail to translate this understanding into accurate predictions (Nearing et al., 2020). However, as our toolkit is founded on both data-driven and theory-driven techniques, it is capable of achieving high predic-

tive performances as with data-driven models while benefiting hydrologists to better understand catchment dynamics through readily interpretable induced models as with theory-driven models. Hence, we used the above sentence to highlight this property of our toolkit.

> L36: see comments above re: data science and “commercial” fields. One can argue that statistical methods have been successfully used in hydrology for decades, and statistics can be considered as a part of “data science”?

[Response] In the context of our manuscript, we refer machine learning models as data science or data-driven models. We will explicitly state this in the revised manuscript.

> “data-driven models are often performing better in terms of predictive capabilities”: I think the authors should provide some references to direct comparisons between physics-based and data-driven hydrological models to support this statement. While I agree that data-driven models can perform exceptionally well in predicting a given output, very little research compares directly with physics-based models. Of course any comparison would also have to include issues related to complexity, computational expense, etc. along with common model performance metrics.

[Response] Yes, we agree there are very little researches which compare data-driven models directly with physics-based models (e.g., Kratzert et al., 2019; Nearing et al., 2018). We also made our statement indirectly based on the following literature and will be added into the revised manuscript.

“It has been the case for a long time that our best process-based hydrology models are less accurate than calibrated conceptual models, which in turn are generally less accurate than even relatively simple data-driven models.” – Nearing et al. (2020)

In the revised manuscript, we will discuss physics-based, conceptual and data-driven models in one section. Hence we will be able to clarify the above statement better in the revised manuscript.

> On a related note: “they may contribute little towards the advancement of scientific discovery due to the lack of interpretability of the model configurations.”: I think the idea of treating data-driven or machine learning methods as a “black box” is not entirely fair. Sufficient information can be gleaned from a simple ANN, as an example, to understand the effects of the forcing on the output, the sensitivity, and so on.

[Response] Yes, we also agree that sufficient information can be gained from simple data-driven models. However, the black-box nature is the most common criticism of data-driven models. We follow the Theory Guided Data Science (TGDS) paradigm (Karpatne et al., 2017) to develop our toolkit. In this TGDS paper, authors find the lack of interpretability of pure machine learning models as one of the major reasons for the limited success of data science models in scientific fields. Addressing this most common criticism on data science models is the motivation behind the emergence of TGDS.

“ The limitations of black-box data science models in scientific disciplines motivate a novel paradigm that uses the unique capability of data science models to automatically learn patterns and models from large data, without ignoring the treasure of accumulated scientific knowledge.” – Karpatne et al. (2017)

“Yet another major limitation of ANNs is in the lack of physical concepts and relations. This has been one of the primary reasons for the skeptical attitude towards this methodology.” – Govindaraju (2000)

> My comments above are not to undermine the central thesis of the paper, I do think there is an interest and a need to study “physics informed machine learning”, but I do not think some of the statements made by the authors are necessary to make this claim. The premise ought to be clarified and perhaps toned down. In addition, I think the authors can cite appropriate literature to better highlight the need for their proposed framework: are their many studies that demonstrate the inability of standard ML approaches to not capture physical phenomenon correctly? Similarly, with respect

to model complexity (objective 2), some reference to current literature that explore complexity for model selection should be highlighted (the “simpler the better paradigm” is not the only paradigm currently being used).

[Response] We believe the importance and necessity of Theory Guided Data Science (Karpatne et al., 2017) or physics informed machine learning are well established among the hydrological modelling community (e.g. Shen, 2018; Nearing et al., 2020). In this study, we also follow the same paradigm and hence address the same issues which the paradigm itself tries to achieve (bring together the existing body of knowledge with machine learning techniques to guide the machine learning algorithms to induce more interpretable models). We arranged the content in the original manuscript to highlight this necessity. However, as you have stated, this might not be as explicit as it should be. Therefore, in the revised manuscript we will explicitly state the need for our framework by citing the relevant literature. The model selection stage of MIKA-SHA aims to identify a model with appropriate complexity (not high, not low) among the competitive models to represent the catchment dynamics (named as the optimal model). We try to achieve this by measuring the model complexity through a combined measure based on model parsimony, information content and pattern matching. Most of such details are presented in the later stage of the original manuscript. Therefore, as you have proposed the role of complexity in model selection will be discussed here in the revised manuscript with appropriate recent literature.

Fundamental Approaches in Hydrological Modelling

> Section 2.1: I don’t think this discussion is really necessary. These concepts have been explored in numerous research papers over the years, and in my opinion are well-understood. The review of current research presented here is short, and no new information or discussion is presented, that isn’t well-understood. I think the manuscript would benefit from reducing length of the text and many of the introductory comments here can be removed, save for presenting the central thesis of TGDS (which to some extent is already presented in the Introduction).

[Response] As you correctly stated, we generally discuss different modelling strategies in this section as some of which are used in our toolkit. For example, between physics-based models and data science models we use data science models (GP), between fixed and flexible conceptual models we use flexible models (FUSE & SUPERFLEX), between lumped and distributed modelling we use distributed modelling (semi-distributed modelling). Therefore, we used a general discussion here as an approach to introduce our novel model induction toolkit. However, as you have proposed we will focus more on TGDS (in the original manuscript we discuss this in the 3rd section) in the revised manuscript.

> “The first reported physics-based model was introduced by Freeze and Harlan (1969).” Perhaps a caveat should be added that this refers to a digitally-simulated hydrological model. Physics based models in general are not that recent – whether a computer was used or not is a different question.

[Response] Yes, we are referring to the first digitally-simulated hydrological model. We will explicitly state this in the revised manuscript.

> “This dichotomy led to the evolution of two major communities in water resources engineering: those who work with physics-based modelling and those who deal with machine learning techniques, which appear to be working quite separately.” – Perhaps this is a generalisation? Many work in both communities simultaneously. This distinction and statement is not necessary for the central thesis of the paper.

[Response] Yes, this is a generalization. As a research group, we also work in both physics-based and data science modelling. Most of the TGDS or physics informed machine learning researchers also work in both communities simultaneously. However, this distinction is identified as one of the major reasons to evolve TGDS as a new modelling paradigm. We made the above statement based on the following literature.

“physical process-oriented modellers have no confidence in the capabilities of data-driven models’ outputs with their heavy dependence on training sets, while the more

system engineering-oriented modellers claim that data-driven models produce better forecasts than complex physically-based models." – Todini (2007)

“[m]any participants who have worked in modeling physical-based systems continue to raise caution about the lack of physical understanding of ML methods that rely on data-driven approaches." - Sellars (2018)

“Regardless of whether hydrologists are willing to accept a strong divergence from what we currently recognize as the body of hydrological theory, it is inevitable that ML will replace large portions of the current focus on process-driven modeling in the next 2-5 years. Simply, ML does the job that hydrologists have failed to do.” – Nearing et al. (2020)

> “The key concept behind this approach is to incorporate the existing body of scientific knowledge into learning algorithms to come up with physically meaningful models with good predictive power.” Repeated text from Introduction.

[Response] Will be removed in the revised manuscript.

> Section 2.2: As per my comment in Section 2.1, much of this discussion will be well known to readers of HESS. Conceptual modelling is well defined and understood. I don’t think it is necessary in this paper.

[Response] We added this section to give a general idea about conceptual models as we are using the conceptual models (specifically flexible modelling frameworks) as the elements of the existing body of hydrological knowledge in the current study. As you have proposed we will reduce the content here and use a single section on physics-based vs. conceptual vs. data science in the revised manuscript.

> L100: equifinality: other reasons for equifinality may also exist than those cited by the authors (e.g., measurement uncertainty, lumping).

[Response] As you correctly said, there are other reasons for equifinality other than parameter uncertainty (e.g. structural uncertainty, measurement uncertainty and lumping). We will also add these reasons in the revised manuscript.

> L119: to “customize model structure” alone cannot be the difference between fixed and flexible?

[Response] Ability to test many hypotheses instead of one fixed hypothesis is the main difference between flexible modelling frameworks and fixed models.

“In contrast to currently dominant “fixed” model applications, the flexible framework proposed here (SUPERFLEX) allows the hydrologist to hypothesize, build, and test different model structures using combinations of generic components.” – Fenicia et al. (2011)

However, this powerful nature of flexible modelling frameworks facilitates hydrologists to use them to address important hydrological issues, such as considering the uniqueness of the place (Beven, 2020), equifinality (Beven, 2012) and identifying model structural errors (e.g. Clark et al., 2008) where the fixed models were unable to do so well.

> “Hence, a hydrologist with novice knowledge would require to test many model structures beforehand selecting an optimal model which is time demanding and computationally intensive, in consequence, hinders the opportunity to use the flexible modelling frameworks in their full potential. Further, the selection of a model configuration without testing a large number of possible combinations may introduce a high level of subjectivity into the model building phase. Therefore, we find a requirement to automate the model building phase to remove the subjectivity and consider many configurations without direct human involvement.” As per my previous comments, much of the preceding text in the manuscript is not necessary, and not properly supported by references to existing literature, to arrive at the goal quoted above. Model selection based on “subjectivity”, or experience, isn’t necessarily bad on its own. I think the authors are simply trying to state model selection and model configuration is an on-going issue in hydrological modelling research (whether physical or data-based), and there is a need to develop a more independent framework to do this.

[Response] As you correctly stated, we also do not see any disadvantage of subjective model selection if the selection is based on the expert’s knowledge and sufficient fieldwork insights about the catchment. However, as we all know this might not be the situation at all time in hydrological modelling. In a recent paper (Addor and Melsen, 2019) based on more than 1500 peer-reviewed research articles, concluded that the model selection in hydrological modelling is more often driven by legacy rather than the adequacy. In such situations, model selection may be governed by the factors, such as the popularity of model, easiness, prior experience instead of the appropriateness of the model for the intended task which may result in biased research findings. However, MIKA-SHA evaluates millions of possible model configurations (hypotheses about catchment behaviour) using the model building components of flexible modelling frameworks (which would be quite impossible to do manually) before selecting an optimal model for the catchment of interest. As you said, we wanted to highlight the importance of developing an independent model induction and selection framework through the content here. We will explicitly highlight this with relevant literature in the revised manuscript.

> Section 2.3: Again, I don’t think the elementary definitions of lumped v distributed v semi-distributed models are necessary. Paper length can be reduced by removing these sections in the review.

[Response] As we mentioned earlier, we use semi-distributed modelling concepts to induce distributed rainfall-runoff models. We added this section to highlight the importance of distributed models over lumped models especially when the catchment size increases and to point out why we selected semi-distributed modelling instead of fully distributed modelling. We will explicitly highlight this in the revised manuscript.

> L176 and 178: citation to appropriate literature are need to support these statements.

[Response] Following references will be added in the revised manuscript to support the statement.

"It is now more than 40 years since Freeze and Harlan published their seminal blueprint for a physically-based digitally-simulated hydrologic response model". "At the time, implementation of those process descriptions was severely limited by the computer power available." – Beven (2012)

"This (applying Darcy–Richards equation for heterogeneous soil), together with the difficulty of knowing the true boundary conditions and characteristics of the subsurface in the field, may certainly be why it has proven rather difficult to show that models based on the Freeze and Harlan blueprint can successfully reproduce the behaviour of real catchments." – Beven (2012)

"Practical issues connected with process-based models, such as difficulty in their use, scalability of physical laws, prohibitive computational times and a large number of parameters, have hampered widespread adoption of these tools." – Fatichi et al. (2016)

"developing a hyper resolution hydrological prediction capability is a grand challenge for hydrology because of the significant modeling, computational, and data needs that will be required for global or continental predictions at these spatial resolutions" – Wood et al. (2011)

> L182: what are the effects of over-parameterisation? I assume the authors are alluding to difficulty in calibration, but this should be explicitly stated.

[Response] Yes, we used the term over-parameterisation to refer to the difficulty in estimation of model coefficients. Too many model coefficients may result in good fitting however may result that the transfer functions may not be physically realistic (Beven, 2012). We will explicitly state this in the revised manuscript.

Machine Learning in Water Resources

> Much of the information presented until Section 3.3 are not really necessary – these are well established within hydrological models. As with previous sections, the current review is not thorough enough (if the intention is to provide a full review), while on the

other hand, not necessary to arrive at the main objective of the paper. Comments below require attention, but my recommendation is to drastically reduce this background information.

[Response] As you proposed, we will revise the first two sections here (Section 3) by reducing the background information. Indeed, our intension was not to provide a thorough review of machine learning applications in water resources. Instead, we guided the interested readers to relevant literature for more details.

> "Another major advantage of a machine learning model is that it requires much less effort to develop and calibrate than a physics-based model." Perhaps a reference to support this statement can be included?

[Response] Following reference will be added in the revised manuscript.

"but the effort needed to build state-of-the-art ML models is orders of magnitude lower than what is required to build process models" – Nearing et al. (2020)

> L216: perhaps replace scientific theories with physical hydrology?

[Response] Yes, we find the term physical hydrology or domain knowledge would be more appropriate here and will be changed in the revised manuscript.

> L247: "several output nodes": should be "one or more" : : :there is not requirement to have more than one node in the output layer.

[Response] Will be corrected in the revised manuscript.

> L253: feed forward is a type, back propagation is a training algorithm.

[Response] Will be corrected in the revised manuscript.

> "ANN can handle incomplete or erroneous data, highly complex and interdependent parameters." I don't agree that "erroneous" data can be "handled" by ANNs per se.

[Response] Erroneous term will be removed in the revised manuscript.

> “One of the key disadvantages of using ANN for data modelling is the fact that it produces overfitting results which make it difficult to extrapolate beyond witnessed train data.” This is contradictory to the statement about ANN handling “noisy data”. Models that are prone to overfitting do not deal well with overfitting. Next, there are several well established methods for preventing overfitting (i.e., regularisation, drop-out, stop training, cross-validation, etc.). Finally, extrapolation on data beyond the training domain is a distinct issue from overfitting and model generalisation. A generalised model should not be expected to perform well on data outside of the training domain.

[Response] Yes, we agree that the ability to handle noise data contradicts with the overfitting issue because overfitting occurs when the network tries to fit the noise component of the data as well. We will correct this in the revised manuscript.

The above statement on overfitting is a generalization of ANN applications in hydrological modelling. We agree with you that there are well-established methods to prevent overfitting. However, discussing such details would be beyond the scope of our manuscript.

“The literature shows that ANNs suffer from some apparent drawbacks and limitations, which are local minima, slow learning speed, over-fitting problem and trivial human intervention such as learning rate, learning epochs and stopping criteria.” – Yaseen et al. (2015)

> “determining the efficient network architecture and tuning hyperparameters make it hard for the user to completely understand how the model makes its predictions.” There is considerable research in this field both in and out of hydrology.

[Response] Again the statement above is a generalization about the ANN applications in hydrological modelling.

“The fact that there is no standardized way of selecting network architecture also receives criticism. The choice of network architecture, training algorithm, and definition

of error are usually determined by the user’s past experience and preference, rather than the physical aspects of the problem.” – Govindaraju (2000)

> L300: “This modelling paradigm aims to simultaneously address the limitations of data science and physics-based models and induce more generalizable and physically consistent models”: This is the major premise of the proposed research. However, I do not think the authors adequately cover existing research efforts in addressing these issues. A fairer comparison would be to simply highlight known issues with ML methods (e.g., “black box”, over fitting, uncertainty, generalisability, model selection) and then highlight current efforts to address these in hydrology. Similarly, known issues with physics-based modelling (e.g., spatial-temporal issues, data resolution, uncertainty, model selection) can be highlighted along with recent efforts to address these issues. Following this, the proposed framework can be described as an alternative way to collectively address these issues. However, the proposed framework is not attempting to address all the limits of both data-centric and physics-based models, rather MIKA-SHA seems to address issues related to spatial heterogeneities and a model selection only.

[Response] In this section, we describe the Theory Guided Data Science paradigm (Karpatne et al., 2017) and not about our toolkit specifically. Yes, we agree that our toolkit can be categorized as a hybrid TGDS approach. Please see the two statements below which have been extracted from the original paper where TGDS was introduced.

“The paradigm of theory-guided data science attempts to address the shortcomings of data-only and theory-only models by seamlessly blending scientific knowledge in data science models” – Karpatne et al. (2017)

“TGDS further attempts to achieve better generalizability than models based purely on data by learning models that are consistent with scientific principles, termed as physically consistent models.” – Karpatne et al. (2017)

Therefore, in the context of TGDS, addressing shortcomings of data-only and theory-only means primarily improving the generalizability and interpretability (pro of theory-only and con of data-only) of data-only models while preserving the high prediction accuracies (pro of data-only and con of theory-only). As a hybrid TGDS approach, MIKA-SHA also tries to achieve the same objectives. Additionally, we try to address the following issues as well.

Avoid overfitting – Limit tree growth, consider validation fitness in the optimal model selection.

Remove subjectivity – Automated process ensures no direct human involvement, test many hypotheses before selecting an optimal model.

Consider the uniqueness of the place – Use of flexible modelling frameworks instead of fixed conceptual models, incorporate spatial heterogeneity of catchment properties and climate variables through semi-distributed modelling.

Handling equifinality – Optimal model selection is based on many absolute and relative performance matrices.

We will explicitly state about the short-comings that we are trying to address through MIKA-SHA in the revised manuscript.

> L313: "ANNs suffer the most severe consequences of lack of interpretability of resulted models" Please provide some evidence for this statement.

[Response] As we have mentioned in the manuscript we limit our discussion on TGDS applications only to ANNs and GP. Hence, the above statement is made relative to the GP. Because while ANN is referred to as a Black-box technique GP is referred to as a Grey-box technique due to its ability to produce explicit mathematical input-output relationships.

"According to the color-based classification of environmental models, GP is more than a common black box data-driven technique. It is a grey box technique that allows the modeler to control model structure, to avoid the typical overfitting problem of traditional

black-box models, and provides better physical interpretation of the process under consideration." – Mehr et al. (2018)

Results

> "Results of this study, such as achieving high efficiency values for the absolute performance measures and obtaining a good visual equivalent between measured and modelled hydrographs suggest that topography of the catchment may have a strong impact on runoff generation." It is not clear to me how the model performance metrics can indicate the latter – please expand.

[Response] Here, the semi-distributed models use topography based HRUs (hills, floodplain, plateau). One can also use either soil type based, geology based, lithology based or even combinations of above as the HRUs. As per the results of the current study, incorporating catchment spatial heterogeneity based on the topography of the catchment is able to capture the runoff dynamics of the catchment reasonably well (high-efficiency values and good visual match between observed and simulated hydrographs). Hence, we may expect the topography of the catchment to be a dominant runoff driver.

> "The consistent performances over the calibration, validation and testing periods of all selected optimal models through MIKA-SHA show no such issues in this case." This is perhaps an over generalisation: (1) Table 8 shows the testing performance is lower for each metric; (2) to conclusively say this, a cross-validation approach is recommended (where the testing dataset is iteratively changed, and the performance under each training scenario is calculated).

[Response] We continuously work on improving our toolkit. Hence, we will consider your recommendation on cross-validation by iteratively changing the training dataset in the next version of MIKA-SHA. However, in the current version of the MIKA-SHA, we have taken the following steps to effectively remove overfitted model configurations.

(1) Once the Pareto-optimal models are identified based on training fitness values (calibration), their fitness values are evaluated on the validation period using the same multi-objective criterion. Then the Pareto-optimal models are reidentified using both calibration and validation fitness values. Through this, toolkit removes the models which perform better only for the calibration period.

(2) Model parsimony based on the number of model parameters is used as a selection criterion in the optimal model selection stage.

(3) Once the optimal model is selected its performance is evaluated on out of sample dataset (testing period).

(4) In the present study, we use the building blocks of two flexible modelling frameworks as the incorporated hydrological knowledge to guide the learning algorithm (as special functions). These model building components themselves follow certain physical laws within their original frameworks (internally coherent). Hence, we expect the models induced using these special functions to be less susceptible to overfitting than models induced using just mathematical functions.

Regarding the testing period performance values, although they are slightly lower than the calibration and validation period performance values (we believe this is the common scenario even in theory-driven models), they still can be considered as relatively high-efficiency values (more than 0.75 where the optimum is at 1) in the hydrological modelling perspective (e.g. It is a common norm that a model with NSE value greater than 0.6 to be considered as a satisfactorily performing model).

> With the results in general, I think the authors can make a stronger attempt to connect to their central thesis. How does MIKE SHA address all the comments and limitations that the authors explored earlier in the paper with respect to data-driven and physics based model. A key limitation, for me, was difficulty in understanding the proposed method in Section 5 (and to some extent in Section 4). Some claims, as commented on above, are difficult to confirm if the results are based on one catchment, under

one training-validation-testing split. The question is if this is repeatable under different conditions?

[Response] We are pleased with your comments. We will consider the suggestions made by you to restructure our paper in the revised manuscript. As per your concern about Section 5 and 4, we will expand those sections by providing more details about our proposed model induction toolkit. Indeed, we have tested MIKA-SHA on many other catchments and it provides equally good results with them. However, in this manuscript, our main objective was to introduce the toolkit rather than focusing more on its applications. We limited the result section into one catchment just to show how MIKA-SHA operates. The unique feature of our toolkit is the readily interpretable nature of induced models. Hence, we allocated some space to demonstrate how induced models can be explained along with catchment characteristics and previous research findings. Therefore, if we to add more catchments, the length of the paper would be increased significantly. However, in our future work, we will focus on applications of MIKA-SHA.

Conclusions

> “may contribute to the development of accurate yet pointless models with severe difficulties with interpretation may not serve towards the advancement of hydrological knowledge” I disagree with the “pointless” claim: : :I think it is OK to highlight some of the limitations of data-driven models, but calling them pointless collectively is not fair.

[Response] Yes, we accept that the term pointless would be misleading. Hence we will remove the term pointless in the revised manuscript.

“may contribute to the development of accurate models with severe difficulties with interpretation may not serve towards the advancement of hydrological knowledge”

Technical corrections

> L122: his/her => their

[Response] Will be corrected in the revised manuscript.

> L211: rewrite for clarity “potential to apprehend the noise complexity”

[Response] Will be rewritten as “ability to capture noise complexity”.

> L242: transpired should be inspired? Not limited to human brains, by the way.

[Response] Yes, “inspired” is more suitable and will be corrected in the revised manuscript.

> L426: “greater extend.” Should be extent?

[Response] Yes, “extent” is the correct word and will be corrected in the revised manuscript.

> L428: “a bunch of”: perhaps “set” is more appropriate than “bunch”

[Response] The term “set” will be used in the revised manuscript.

> Table 4: “ln” is defined at the bottom of the table, but the equation uses “log”; unsure if the log is ln or base10. Please clarify.

[Response] Yes, the base in the equation is “e” (Natural logarithm). logNSE is the short name for log Nash-Sutcliffe efficiency.

REFERENCES

Addor, N., and Melsen, L. A.: Legacy, rather than adequacy, drives the selection of hydrological models. Water Resources Research, 55. https://doi.org/10.1029/2018WR022958, 2019.

Beven, K.: Rainfall-runoff modelling: the primer, Wiley-Blackwell, West Sussex, United Kingdom, 2012.

Beven, K.: Deep Learning, Hydrological Processes and the Uniqueness of Place, Hydrological Processes, in press, https://doi.org/10.1002/hyp.13805, 2020.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., et al.: Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, Water Resources Research, 44, W00B02, https:// doi.org/10.1029/2007WR006735, 2008.

Fatichi, S., Vivoni, E. R., Ogden, F. L., Ivanov, V. Y., Mirus, B., Gochis, D., et al.: An overview of current applications, challenges, and future trends in distributed process-based models in hydrology, Journal of Hydrology, 537, 45-60, doi:10.1016/j.jhydrol.2016.03.026, 2016.

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

Govindaraju, R. S.: Artificial neural networks in hydrology. II: Hydrologic applications, Journal of Hydrologic Engineering, 5, 124–137, 2000.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., et al.: Theory-guided data science: A new paradigm for scientific discovery from data, IEEE Transactions on Knowledge and Data Engineering, 29, 2318–2331, doi:10.1109/TKDE.2017.2720168, 2017.

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., and Nearing, G.: Toward improved predictions in ungauged basins: Exploiting the power of machine learning, Water Resources Research, 55, 11344–11354, https://doi.org/10.1029/2019WR026065, 2019.

Mehr, A. D., Nourani, V., Kahya, E., Hrnjica, B., Sattar, A. M. A., and Yaseen, Z. M.: Genetic programming in water resources engineering: A state-of-the-art review, Journal of Hydrology, 566, 643–667, 2018.

Nearing, G., Yatheendradas, S., Crow, W., Zhan, X., Liu, J., and Chen, F.: The efficiency of data assimilation, Water resources research, 54 (9), 6374-6392, 2018.

Nearing, G. S., Kratzert, F., Sampson, A. K., Pelissier, C. S., Klotz, D., Frame, J. M., and Gupta, H. V.: What Role Does Hydrological Science Play in the Age of Machine Learning?, Water Resources Research, in press, 2020.

Sellars, S.: “grand challenges" in big data and the earth sciences. Bulletin of the American Meteorological Society, 99 (6), ES95-ES98, 2018.

Shen, C.: A trans-disciplinary review of deep learning research and its relevance for water resources scientists, Water Resour. Res., https://doi.org/10.1029/2018WR022643, 2018.

Todini, E.: Hydrological catchment modelling: past, present and future, Hydrology and Earth System Sciences, 11 (1), 468-482, 2007.

Wood, E. F., Roundy, J. K., Troy, T. J., Van Beek, L., Bierkens, M. F., Blyth, E., . . .et al.: Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring earth’s terrestrial water. Water Resources Research, 47 (5), 2011.

Yaseen, Z.M., El-shafie, A., Jaafar, O., and Afan, H.A.: Artificial intelligence based models for stream-flow forecasting: 2000-2015, Journal of Hydrology, 530, 829-844, http://dx.doi.org/10.1016/j.jhydrol.2015.10.038, 2015.

Interactive comment on Hydrol. Earth Syst. Sci. Discuss., https://doi.org/10.5194/hess-2020-487, 2020.