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 reviewers' comment is marked by starting the line with "[Comment]", while the corresponding response is annotated with "[Response]". The changes made in the revised manuscript are given by starting the line with "[Changes]". The line numbers in reviewers' comments refer to the original manuscript while the line (L) and page numbers (P) in the author's response refer to the revised manuscript. In this document, all reviewers' comments are numbered (e.g. reviewer 1's comments start with 1.1 and reviewer 2's comments start with 2.1). In addition to this document, a marked-up revised manuscript is provided where the changes are highlighted and a note is added to each change referring to the relevant reviewers' comment (e.g. As per the reviewer comment 1.1).

Reviewer 1:

GENERAL COMMENT

1.1

[Comment] There are many extremely long paragraphs (e.g. lines 70 to 90) that express multiple concepts. I would make paragraphs shorter, creating a paragraph per concept. This should help the readability of the paper.

[Response] We accept that there are many long paragraphs which can be divided into small paragraphs to improve the readability of the paper.

[Changes] Long paragraphs in the original manuscript were divided into short paragraphs in the revised manuscript.

1.2

[Comment] There are several statements without a reference justifying them.

[Response] The relevant references are added appropriately in the revised manuscript

[Changes] Following references were added in the revised manuscript.

- Fenicia et al. (2011) (L128, P5).
- Kratzert et al. (2018, 2019a, 2019b); Nearing et al. (2020b) (L267, P9).
- Knoben et al. (2020) (L131, P5).
- Beven (2012c) (L172, P6).

OVERALL STRUCTURE

1.3

[Comment] I do not like the structure of the paper and the amount of content given to each paragraph. Your main message is to present MIKA-SHA but this is left to section 4, which is barely 1 page out of 42. If that was the main concept of the paper, I would give it more space.

I like the material in sections 1 to 3 (included) but they are basically a mix of introduction and a "methods" section. I would move some content from section 2 and 3 to the introduction, which can be divided in subsections. A possible structure of the introduction can be:

- Quick introduction on hydrological modelling and TGDS

- On hydrological models

* Physics-based models vs. conceptual models vs. data science models (mix of 2.1 and 2.2)

* Focusing on conceptual models, difference between fix and flexible structure (some part of 2.2)

* Lumped vs. distributed (2.3)

- On ML models

* Some generalities (3)

* ANN (3.1, maybe reduce)

* GP (3.2)

- Physics informed ML (3.3)

I would then add a "methods" part where you write about what of sections 2 and 3 you actually use: SUPERFLEX, FUSE, GP. I would also add to this the metrics that you later use in section 5.

Section 5 and 6 are about the case study and I think this should represent a minor part of the paper (which objective is to present how MIKA-SHA works). To this end, I would more or less keep the same content and put it in a single section divided in:

- Presentation of the case study

- Settings

- Results that you get

- Meaningful discussion, potentially showing that MIKA-SHA works. I would therefore move some aspects of this section elsewhere:
- Metrics to a “methods” section
- Further explanation of MIKA-SHA functioning to the section that presents the model
- General implications on the goodness of the approach to a general “discussion and conclusion” section.

[Response] We find the structure that you have proposed is more organised than the original order. Hence, the proposed structure has been used to organise the content in the revised manuscript. The section on MIKA-SHA has been expanded by providing more details in the revised manuscript (details are given below). A single section on results (combining Sect. 5 and Sect. 6 in the original manuscript) is added as proposed here in the revised manuscript. Further, some parts of the results section in the original paper are moved to other sections as proposed.

[Changes] The revised manuscript is organised under following headings.

1 Introduction (L29, P2 – L73, P3)

2 Fundamental Approaches in Hydrological Modelling (L74, P3 – L198, P7)

    2.1 Physics-based Models vs. Conceptual Models vs. Data Science Models (L77, P3 – L117, P4)

    2.2 Fixed Models vs. Flexible Models (L118, P5 – L151, P6)

    2.3 Lumped Models vs. Distributed Models (L152, P6 – L198, P7)

3 Machine Learning in Water Resources (L199, P7 – L293, P10)

    3.1 Artificial Neural Networks (ANN) (L245, P9 – L269, P9)

    3.2 Genetic Programming (GP) (L270, P9 – L293, P10)

4 Physics Informed Machine Learning (L294, P10 – L343, P12)

5 Methods (L344, P12 – L438, P15)

    5.1 SUPERFLEX (L370, P13 – L378, P13)

    5.2 FUSE (L379, P13 – L386, P13)

    5.3 Performance Measures (L387, P13 – L438, P15)

6 MIKA-SHA (L440, P16 – L554, P20)

    6.1 Data Preprocessing (L471, P18 – L477, P18)

    6.2 Model Identification (L478, P18 – L506, P19)

    6.3 Model Selection (L507, P19 – L521, P19)

    6.4 Uncertainty Analysis (L522, P19 – L554, P20)

7 Application of MIKA-SHA (L555, P20 – L700, P30)

    7.1 Study Area (L559, P20 – L592, P23)

    7.2 Results (L593, P24 – L700, P30)

7.2.1 MIKA-SHA Models induced using SUPERFLEX Building Blocks (L594, P24 – L652, P27)

7.2.2 MIKA-SHA Models induced using FUSE Building Blocks (L653, P27 – L700, P30)

8 Discussion and Conclusions (L701, P30 – L754, P32)

Appendix A (L755, P32 – L766, P35)

References (L778, P36 – L1086, P45)

The changes made here can be summarised as follows,

- The discussion on conceptual models was included along with physics-based and data science models (L88 – L98, P4).
- The discussions on SUPERFLEX and FUSE were moved to the Methods section (L370 – L386, P13).
- Shortened the content on ANNs (more details given below) (L245 – L269, P9).
- Content on physic informed machine learning was moved to a separate section (Section 4) (L294, P10 – L343, P12).
- A new section was introduced as Methods (L344, P12 – L438, P15).
- Details about performance measures used in the study were moved to the Methods section (L387, P13 – L438, P15).
- The section on MIKA-SHA was expanded by including the details of the four major steps of MIKA-SHA (L440, P16 – L554, P20).
- The results of the study were included as a subsection of Applications of MIKA-SHA (L593, P24 – L700, P30).
- General findings of the research were included in the Discussion and Conclusions section (L702, P30 – L732, P31).

PAPER CONTENT

1.4

[Comment] Keeping in mind that I do not have a deep knowledge of GP, I find quite difficult to understand what MIKA-SHA actually does and my difficulty can be motivated by the following reasons:

- Use of jargon from GP, that may be not common in the hydrological community (e.g. model induction vs. model selection)
- Assuming good familiarity of the reader with GP
- Assuming that the reader knows ML-RR-MI: you do not have to re-write here that paper but at least explain here the concepts that are necessary. I find it difficult to follow something that says that MIKA-SHA is basically ML-RR-MI plus something else.

[Response] We accept the concern you made here. Hence, in the revised manuscript, a separate paragraph about the ML-RR-MI is added when MIKA-SHA is introduced. Further, the workflow of the MIKA-SHA has been explained in a more general way of avoiding jargon specific terms.

[Changes] Following changes were made in the revised manuscript.

- Briefly explained about our prior model induction toolkit "ML-RR-MI" (Chadalawada et al., 2020) in Sect. 6 (L441-L452, P16).
- The workflow of MIKA-SHA especially the section on GP based optimization (Model Identification) is explained in a simpler manner (L478, P18 – L506, P19).

DATA AND CODE AVAILABILITY

1.5

[Comment] I do not know if HESS forces the sharing of the source code but I believe that, potentially, MIKA-SHA can be a valuable tool for the hydrological community and, therefore, I invite you to make it publically available.

[Response] We are pleased with your comment. Currently, we are adding several other building blocks to the toolkit where the hydrologists may have more flexibility when they use MIKA-SHA. We are planning to make the toolkit public this year (2021) with a user manual. Further, we need to gain permissions from some other authors before making MIKA-SHA public because we are using their findings as the elements of hydrological knowledge incorporated within the toolkit.

FURTHER COMMENTS

1.6

[Comment] L18: MIKA-SHA - do you want to make it look like MIKE-SHE? If yes, motivate; if no, keep this in mind

[Response] Yes, our name is inspired by the popular distributed hydrological model MIKE-SHE.

1.7

[Comment] L19: meaning of induces? "creates"??

[Response] Yes, it means creates or generates.

1.8

[Comment] L21: What do you mean with "internally coherent"

[Response] We use the term "internally coherent collection of building blocks" to refer the building blocks themselves connect or follow in a reasonable way and each part of them are carefully considered within the framework. A good example of this would be the model-building components of a flexible modelling framework.

1.9

[Comment] L27: "signatures" has a specific meaning in hydrology, which may be different from your intentions

[Changes] Term “signatures” is changed to “dynamics” in the revised manuscript (L31, P2).

1.10

[Comment] L30: hydrological model - hydrological model structure

[Changes] Corrected in the revised manuscript (L34, P2).

1.11

[Comment] L44: induction - consider changing with "selection" since more common in hydrology. Here and afterwards

[Response] In this manuscript, we specifically chose the term induction instead of selection to state that the algorithm itself builds the model structure within its optimization process rather than selecting an already available model structure. As you pointed out the term selection would be more common in hydrology however, the term induction is frequently used in GP applications in hydrology.

1.12

[Comment] L65: vs. (lower case and with the dot at the end) (here and elsewhere)

[Changes] Corrected in the revised manuscript (L77, P3; L118, P5; L152, P6).

1.13

[Comment] L66: this definition applies both to physical based (in the sense of models that are intended to represent the phenomena happening in the catchment) and conceptual models (reservoir models). Clarify better what you mean.

[Response] Yes, this definition applies to the conceptual models as well. Since this section in the original manuscript differentiates physics-based models and data science models, we used this definition only for physics-based models.

[Changes] As we have used the structure you proposed for the revised manuscript where the section heading is physics-based models vs. conceptual models vs. data science models, we have used the same definition on physics-based and conceptual models (L78 – L80, P3).

1.14

[Comment] L78-79: elaborate more. also on the role of conceptual models. In theory if a physics-based model has only pars that can be related to physical measurable properties then you have a different data requirement compared to conceptual models that calibrate their parameters based on model outputs.

[Response] As you correctly said, in this context, physics-based models and conceptual models may have different data requirements (e.g. physics-based models require measured physical quantities to use as model parameter values while conceptual models need measured catchment responses to calibrate the model parameters).

[Changes] The differences in data requirements between the conceptual and physics-based models are elaborated in the revised manuscript (L104 – L107, P4).

1.15

[Comment] L86: was – has?

[Changes] Corrected in the revised manuscript (L115, P4).

1.16

[Comment] L93: I would consider introducing conceptual models in the previous paragraph, in contrast with physically based and data driven. Here you can narrow the focus on conceptual and bring the differentiation between fixed and flexible

[Changes] As proposed the general description on conceptual models was moved to the Sect. 2.1 in the revised manuscript (L88 – L98, P4).

1.17

[Comment] L111-113: bring references to this statement

[Changes] Following reference is added in the revised manuscript (L128, P5).

"A simple illustration of the practical limitations of a fixed model structure is the need to add specialized modules for specific catchment conditions. For example, in many models, the simulation of snowmelt requires the addition of an external snow module, already implying that the overall model structure requires customization for a specific climatic region." – Fenicia et al. (2011)

1.18

[Comment] L115-116: (Knoben, 2019) - paper 2020 wrr on applying the models on the CAMELS

[Changes] The mentioned recent paper (Knoben et al., 2020) has also been cited in the revised manuscript (L131, P5).

1.19

[Comment] L130: subjectivity - motivate more on this. usually the best model solution is selected testing different options which, while being subjective, are designed to cover a wide range of possibilities, in order to exclude subjectivity from the process.

[Response] We use the term subjectivity here to refer the model structure selection or development based on personal judgement, preference and experience. We recognise that the model selection or development based on expert's knowledge can be as good as those through an automative process (can even be better than models through an automative process). However, expert knowledge might not be available and may be expensive. In such situations, we believe an automative model building algorithm would be more appropriate.

[Changes] More elaboration on subjectivity is added in the revised manuscript (L146, P5 – L150, P6).

1.20

[Comment] L169-171: maybe a ref on this

[Changes] Following reference is added in the revised manuscript.

"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. The expansion of computer power in the last 40 years has, however, greatly relaxed this constraint and it is possible now to apply such models with a fine discretisation to both small and large catchments." – Beven (2012c) (L172, P6).

1.21

[Comment] L177: will – would?

[Changes] Corrected in the revised manuscript (L179, P6).

1.22

[Comment] L205-206: kratzert 2017 hess and subsequent papers

[Response] We appreciate pointing out the related work of Kratzert.

[Changes] Related citations (Kratzert et al., 2018; 2019a, 2019b) are added to the revised manuscript (L209 – L210, P7).

1.23

[Comment] L213: process-based - keep consistency. it was physics based before

[Response] Corrected in the revised manuscript (L216, P8).

1.24

[Comment] L214: much less effort - depends how you define effort. human effort -> yes computational effort -> no (probably)

[Response] Yes, as you said it depends on how we define it. We used the term to refer to human effort.

[Changes] Term "human" is added before the term "effort" in the revised manuscript (L217, P8).

1.25

[Comment] L216: scientific theories - maybe better domain knowledge, although data science without domain knowledge is the perfect recipe for a disaster

[Response] Yes, the term "domain knowledge" would be the more suitable term here.

[Changes] Term "domain knowledge" is used in the revised manuscript (L220, P8).

1.26

[Comment] L231: which – that?

[Changes] Corrected in the revised manuscript (L235, P8).

1.27

[Comment] L241: Overall, I don't like the structure. I would write sth like:

- AAN is a subset of ML
- describe briefly AAN and their applications in hydrology
- move to RNN and LSTM, explaining why they are useful to model RR

[Changes] The section on ANN has been restructured as you proposed here in the revised manuscript (L245 – L269, P9).

1.28

[Comment] L255-261: provide some references for the points you are addressing here

[Response] In the original paper we made these statements based on the following literature.

"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)

"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)

[Changes] In the revised manuscript we have removed these sentences and limited the discussion on ANN only to the sections that you have proposed in your previous comment (1.27).

1.29

[Comment] L272: you are missing all the kratzert-nearing part

[Response] We find the related work of Kratzert and Nearing are highly appropriate for the content of our paper. Hence, the following citations are added in the revised manuscript.

[Changes] Rainfall-runoff modelling: Kratzert et al. (2018, 2019a, 2019b); Nearing et al. (2020b) (L267, P9).

1.30

[Comment] L286: It is not clear what differentiates GP from ANN. To my understanding, it looks like ANN have predefined functions and you tune the parameters in training while GP writes the functions.

[Response] Yes, your understanding is correct. In the context of this application, there is no need to predefine the model structure for GP. GP itself develops (induce) the model structure using the available building blocks. GP is capable of optimizing both model structure and parameters simultaneously. Further, the capability to produce explicit mathematical input-output relationships makes GP distinctly different from the rest of the machine learning techniques.

1.31

[Comment] L326: Declarative bias and preferential bias - explain this in simple words. The audience may not know GP

[Response] The two terms are explained in simple words in the revised manuscript. In the related work (Keijzer and Babovic, 2002), authors have introduced declarative and preferential bias into the GP algorithm to induce dimensionally correct equations. At the initialization stage, declarative bias forces to sample only the dimensionally correct solutions (a hard constraint on dimensional correctness). On the other hand, preferential bias guides the algorithm towards the dimensionally correct solution (a soft constraint on dimensional correctness) while allowing all solutions to induce.

[Changes] The above explanation is added in the revised manuscript (L325 – L328, P11).

1.32

[Comment] L344: I would make a separate section out of this

[Changes] As proposed a separate section on the content here is added under the Methods section in the revised manuscript (L345, P12 – L369, P13).

1.33

[Comment] L364: I would extend this section a lot, maybe including elements of the previous sections. Really difficult to read for somebody that does not know GP. Try to explain with less technical jargon. Plus you are explaining the model like "it is ML-RR-MI + sth" which requires to read about ML-RR-MI.

[Response] As you have stated here and in some earlier comments, we have expanded this section (where MIKA-SHA is introduced) by including all the relevant parts from other sections. As mentioned above, we have included a paragraph on how GP based optimization framework works more simply in the revised manuscript. Further, we have added a paragraph explaining the basics of our prior toolkit ML-RR-MI in this section.

[Changes] Following changes have made in this section.

- A paragraph on ML-RR-MI is added (L441 – L453, P16).

- Details about the four major stages of MIKA-SHA are included under this section (L471, P18 – L544, P20).
- GP based optimization framework (Model Identification stage) is explained with less technical jargon (L479, P18 – L506, P19).

1.34

[Comment] L381: multi-objective optimization scheme - did you make a package? Public?

[Response] We have used the approach proposed in NSGA-II (Deb et al., 2002) as the multi-objective optimization scheme in our toolkit. NSGA-II is an augmented version of the popular Genetic Algorithm (GA) which facilitates multi-objective optimization based on Pareto-optimality concept. NSGA-II is a publicly available package for GA. However, the evolution process in both GA and GP are much similar. Hence, we developed our optimization scheme based on the concepts in NSGA-II.

1.35

[Comment] L381: NSGA-II - overall, the procedure is not very clear to me. Maybe it requires large knowledge of GP or the other framework but I don't think you should rely on it.

[Response] As mentioned in the earlier response, NSGA-II is a well established multi-objective optimisation method not only in GP but also in other evolutionary computation techniques like Genetic Algorithm (GA). The approach is based on the Pareto-optimality concept. As NSGA-II has been used extensively in many previous studies and the package is publicly available, we avoided explaining it deeper in our paper. However, in our revised manuscript we have briefly explained how our multi-objective optimization scheme operates based on NSGA-II.

[Changes] Details about the multi-objective optimization scheme of MIKA-SHA are added in the revised manuscript (L494, P18 – L506, P19).

1.36

[Comment] L433: objective functions - this should go into a methods section.

[Changes] Moved to the Methods section in the revised manuscript (Table 1, P14).

1.37

[Comment] L450: SIS - shouldn't this go into methods?

[Changes] Moved to the Methods section in the revised manuscript (L408, P14 – L420, P15).

1.38

[Comment] L464: 4. - is this a continuation of the previous list?

[Response] Yes, it is a continuation.

[Changes] As the explanation on SIS is moved to the Method section in the revised manuscript, now it is clear that content here is a complete list (L510 – L521, P19).

1.39

[Comment] L520: Results - if sect 5 is on the case study, shouldn't this be a subsection of 5?

[Response] Yes, it would be more appropriate as a subsection of the case study section (Application of MIKA-SHA).

[Changes] The results section is moved as a subsection (Sect. 7.2) of the case study section (Sect. 7) in the revised manuscript (L593, P24 – L700, P30).

1.40

[Comment] L522: induce - use of the word "induce" unclear

[Response] As earlier the term induce is used to refer create, develop, or generate.

1.41

[Comment] L563: This look like a discussion of the model results, which is nice to have. I would make it more systematic comparing what the model says to your hydrological knowledge (e.g. the model suggests this structure for this HRU and we believe is correct because **) Try to give it a structure..not a single paragraph...

[Response] As you proposed earlier, we have moved the general discussion of MIKA-SHA's optimal models here to a separate section under "Discussion and Conclusions" in the revised manuscript (L702, P30 – L731, P31).

1.42

[Comment] L576: 34 model parameters - it's a lot. I'm surprised it does not overfit

[Response] As our toolkit uses semi-distributed modelling concepts to induce distributed rainfall-runoff models, it assigns a separate model structure to each HRU which may result in high model parameters. However, we use following steps within our framework to remove overfitted models.

- 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.
- Model parsimony based on the number of model parameters is used as a selection criterion in the optimal model selection stage.
- Once the optimal model is selected its performance is evaluated on out of sample dataset (testing period).
- 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.

[Changes] Methods used to avoid overfitting are explicitly stated in the revised manuscript (L549 – L551, P20).

1.43

[Comment] Figure 8: Floodplain - not clear if the division in hill/floodplain/plateau is done by the model or predefined by the user

[Response] The division is predefined by the user. The model assigns sperate model structure to each HRU and calculates the total runoff as per semi-distributed rainfall-runoff paradigm.

1.44

[Comment] L623: as before, make more systematic

[Changes] As mentioned above we have moved the general discussions here to a separate section (Discussion and Conclusions) in the revised manuscript (L702, P30 – L731, P31).

1.45

[Comment] L632: since there are similarities (and probably differences) between FUSE-TOPO-M1 and SUPERFLEX-TOPO-M2 it would be nice to have a paragraph analyzing this. Having a consistent, but slightly different, model structure selection between the two configurations would be a strong point in favor of MIKA-SHA

[Response] Indeed this is a strong characteristic in favour of MIKA-SHA toolkit. Hence, we have added a separate paragraph describing the similarities between the two optimal models derived using the model components of two libraries in a separate paragraph in the Discussion and Conclusions section in the revised manuscript.

[Changes] A paragraph highlighting the similarities of the two optimal models is added in the Discussion and Conclusions section (L720 – L725, P31).

1.46

[Comment] L634: this looks like conclusions of the case study. Re-consider the division in sections.

[Changes] Moved to the Discussion and Conclusion section in the revised manuscript (L723 – L725, P31).

1.47

[Comment] L685: Data availability - what about the code of the model?

[Response] As answered earlier (1.5), we are working on making the MIKA-SHA code public.

Reviewer 2:

GENERAL COMMENTS

2.1

[Comment] 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".

[Changes] Corrected in the revised manuscript (L17 – L18, P1).

2.2

[Comment] 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. We have explicitly stated this in the revised manuscript.

As you have suggested, we have expanded the methods sections by providing more details about our proposed toolkit in the revised manuscript. Regarding the content in the first 3 sections of the original manuscript, we have organised more or less the same content (because Reviewer 1 finds them as meaningful) in a different order in the revised manuscript. The focus of the MIKA-SHA is 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.

[Changes] Following changes were made to address the concerns raised here.

- Explicitly stated the focus of the general discussion on introductory sections (L75 – L76, P3; L206 – L207, P7).
- A new section is added as “Methods” where the specific modelling strategies (Genetic Programming, flexible modelling frameworks and semi-distributed modelling) used within MIKA-SHA is highlighted (L344, P12 – L439, P15).
- The section on MIKA-SHA is expanded by adding details about all four stages of the framework (L440, P16 – L554, P20).
- The focus of MIKA-SHA is explicitly stated in the revised manuscript (L60 – L61, P3; L545 – L554, P20).

SPECIFIC COMMENTS

Abstract

2.3

[Comment] 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. (2020a).

“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. (2020a).

“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).

2.4

[Comment] 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.

[Changes] As it would be confusing to the readers, the two frameworks are introduced in the introduction section and are removed from the abstract in the revised manuscript (L57 – L59, P2).

2.5

[Comment] 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”.

[Changes] Added 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 (L22 – L26, P1).

Introduction

2.6

[Comment] “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] We 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.

[Changes] The term “signatures” is replaced with term “dynamics” (L31, P2).

2.7

[Comment] “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 (2012a), “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 (2012a).

“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).

[Changes] Above references are added to support the statement (L35, P2).

2.8

[Comment] “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 (2012c).

“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 (2012c).

[Changes] Above references are added to support the statement (L36, P2).

2.9

[Comment] “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., 2020a). However, as our toolkit is founded on both data-driven and theory-driven techniques, it is capable of achieving high predictive 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.

2.10

[Comment] 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.

[Changes] We have stated this in the revised manuscript (L41, P2).

2.11

[Comment] “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 that there are very little studies which compare data-driven models directly with physics-based models (e.g., Nearing et al., 2018; Kratzert et al., 2019a). We also made our statement based on the following literature.

“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. (2020a)

[Changes] Related references are added (L44, P2).

2.12

[Comment] 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)

[Changes] A reference is added (Karpatne et al., 2017) to support the statement (L45, P2).

2.13

[Comment] 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., 2020a). 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. Hence we have explicitly stated the target of our proposed framework in the revised manuscript.

As you have correctly said the “simpler the better” paradigm is not the only paradigm use to select models. What we really wanted to state here was that we have employed a quantitative model selection approach in MIKA-SHA instead of the “simpler the better” paradigm used in our prior toolkit (ML-RR-MI). We have clearly stated this in the revised manuscript. 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. However, the measures used to assess the model complexity are presented in the Methods section in the revised manuscript.

[Changes] Following changes are made in the revised manuscript to address the concerns here.

- The target of MIKA-SHA is explicitly stated in the revised manuscript (L60 – L61, P3; L545 – L554, P20).
- Clearly stated that a new quantitative model selection approach is used in MIKA-SHA as opposed to the “simpler the better” paradigm used in ML-RR-MI (L63 – L64, P3).
- Complexity measures used in MIKA-SHA are summarized in the Methods section (L399, P14 – L439, P15).

Fundamental Approaches in Hydrological Modelling

2.14

[Comment] 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.

[Changes] Specific modelling strategies used in the proposed toolkit (out of the modelling strategies presented in Sect. 2) are discussed in a separate section under Methods section in the revised manuscript (L345, P12 – L386, P13).

2.15

[Comment] "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.

[Changes] Term "digitally-simulated" is added in the revised manuscript (L82, P3).

2.16

[Comment] "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. (2020a)

[Changes] Above references are added to support the statement (L114, P4).

2.17

[Comment] “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.

[Changes] Removed the repetition in the revised manuscript.

2.18

[Comment] 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.

[Changes] We have used a single section on physics-based models vs. conceptual models vs. data science models in the revised manuscript (L77, P3).

2.19

[Comment] 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).

[Changes] We have also added the other reasons for equifinality in the revised manuscript (L96, P4).

2.20

[Comment] 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, 2012a) and identifying model structural errors (e.g. Clark et al., 2008) where the fixed models were unable to do so well.

2.21

[Comment] "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.

[Changes] Explicitly highlighted the importance of an independent model induction and selection framework with relevant literature in the revised manuscript (L146, P5 – L151, P6).

2.22

[Comment] 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.

[Changes] Explicitly highlighted the reasons for selecting the semi-distributed modelling paradigm in the revised manuscript (L345 – L349, P12).

2.23

[Comment] L176 and 178: citation to appropriate literature are need to support these statements.

[Response] We made the statement based on the following literature.

"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 (2012c)

“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 (2012c)

“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)

“One approach to the problem is to try and build in what information we do have into a distributed model of the processes. As Nearing et al. (2020) point out, however, there have been millions of dollars invested in such models without real demonstration that this approach is successful (this was already the case in 1987 but much more has been spent since).” – Beven (2020)

[Changes] The above references are added in the revised manuscript (L178, P6; L180, P6).

2.24

[Comment] 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, 2012b).

[Changes] The effects of over-parameterisation are added in the revised manuscript (L184 – L185, P7).

Machine Learning in Water Resources

2.25

[Comment] 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] 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. We wanted to highlight the importance and potential of data science models in hydrological modelling while pointing out their limitations as some of the readers may not be much familiarized about machine learning.

[Changes] Most of the background information on ANN are removed in the revised manuscript (L245 – L269, P9).

2.26

[Comment] “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?

[Changes] Following reference is added in the revised manuscript (L218, P8).

“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. (2020a)

2.27

[Comment] L216: perhaps replace scientific theories with physical hydrology?

[Response] Yes, we find the term physical hydrology or domain knowledge would be more appropriate here.

[Changes] Term “domain knowledge” is used in the revised manuscript (L220, P8).

2.28

[Comment] L247: “several output nodes”: should be “one or more” : : :there is not requirement to have more than one node in the output layer.

[Changes] Corrected in the revised manuscript (L250 – L251, P9).

2.29

[Comment] L253: feed forward is a type, back propagation is a training algorithm.

[Changes] Removed the sentence in the revised manuscript.

2.30

[Comment] "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.

[Changes] Removed the sentence in the revised manuscript.

2.31

[Comment] "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.

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)

[Changes] Removed the sentence in the revised manuscript.

2.32

[Comment] "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)

[Changes] Removed the sentence in the revised manuscript.

2.33

[Comment] 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.

[Changes] We have explicitly stated about the short-comings that we are trying to address through MIKA-SHA in the revised manuscript (L60 – L64, P3; L545 – L554, P20).

2.34

[Comment] 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)

[Changes] Explicitly stated that the above statement on ANN is made relative to the GP and the given reference is added in the revised manuscript (L311, P11).

Results

2.35

[Comment] “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 these 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.

[Changes] Explicitly stated why we may expect topography to have a strong impact on runoff generation in Red Creek catchment (L702, P30 – L709, P31).

2.36

[Comment] "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.

- 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.
- Model parsimony based on the number of model parameters is used as a selection criterion in the optimal model selection stage.
- Once the optimal model is selected its performance is evaluated on out of sample dataset (testing period).
- 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)

[Changes] Techniques used to avoid overfitting are explicitly stated in the revised manuscript (L549 – L551, P20).

2.37

[Comment] 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 have considered the suggestions made by you to restructure our paper in the revised manuscript. As per your concern about Section 5 and 4, we have expanded 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

2.38

[Comment] "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.

[Changes] Removed the term "pointless" in the revised manuscript (L745 – L747, P32).

Technical corrections

2.39

[Comment] L122: his/her => their

[Changes] Corrected in the revised manuscript (L137, P5).

2.40

[Comment] L211: rewrite for clarity “potential to apprehend the noise complexity”

[Changes] Rewritten as “ability to capture noise complexity” (L214, P8).

2.41

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

[Changes] Corrected in the revised manuscript (L247, P9).

2.42

[Comment] L426: “greater extend.” Should be extent?

[Changes] Corrected in the revised manuscript (L491, P18).

2.43

[Comment] L428: “a bunch of”: perhaps “set” is more appropriate than “bunch”

[Changes] The term “set” is used in the revised manuscript (L491, P18).

2.44

> 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.

[Changes] Used the term “log” throughout the equation and removed the term “ln” (Table 1, P14).

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.: Down to basics: Runoff processes and the modelling process, in: Rainfall-runoff modelling: the primer, Wiley-Blackwell, West Sussex, United Kingdom, 1–22, 2012a.

Beven, K.: Predicting Hydrographs Using Models Based on Data in: Rainfall-runoff modelling: the primer, Wiley-Blackwell, West Sussex, United Kingdom, 83–118, 2012b.

Beven, K.: Beyond the Primer: Next Generation Hydrological Models, in: Rainfall-runoff modelling: the primer, Wiley-Blackwell, West Sussex, United Kingdom, 313–327, 2012c.

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

Chadalawada, J., Herath, H. M. V. V., and Babovic, V.: Hydrologically informed machine learning for rainfall-runoff modeling: A genetic programming-based toolkit for automatic model induction, Water Resources Research, 56, https://doi.org/10.1029/2019WR026933, 2020.

Clark, M. P., Slater, A. G., Rupp, D. E., Woods, R. A., Vrugt, J. A., Gupta, H. V., Wagener, T., and Hay, L. E.: 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., Downer, C. W., Camporese, M., Davison, J. H., Brian A. Ebel, B. A., Jones, N., Kim, J., Mascaro, G., Richard G. Niswonger, R. G., Restrepo, P., Rigon, R., Shen, C., Sulis, M., and David Tarboton, D.: 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.

Fenicia, F., Kavetski, D., Savenije, H. H., and Pfister, L.: From spatially variable streamflow to distributed hydrological models: Analysis of key modelling decisions, Water Resources Research, 52, 954–989, doi:10.1002/2015WR017398, 2016.

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., Shekhar, S., Samatova, N., and Kumar, V.: 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.

Kavetski, D., and Fenicia, F.: Elements of a flexible approach for conceptual hydrological modeling: 2. Application and experimental insights, Water Resources Research, 47, W11511, https://doi.org/10.1029/2011WR010748, 2011.

Keijzer, M., and Babovic, V.: Declarative and preferential bias in GP-based scientific discovery, Genetic Programming and Evolvable Machines, 3, 41–79, 2002.

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

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.: Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks, Hydrol. Earth Syst. Sci., 22, 6005–6022, https://doi.org/10.5194/hess-22-6005-2018, 2018.

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, 2019a.

Kratzert, F., Klotz, D., Schulz, K., Klambauer, G., Hochreiter, S., and Nearing, G.: Benchmarking a Catchment-Aware Long Short-Term Memory Network (LSTM) for Large-Scale Hydrological Modeling, Hydrol. Earth Syst. Sci. Dis., https://doi.org/10.5194/hess-2019-368, 2019b.

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, doi: 10.1029/2020WR028091, in press, 2020a.

Nearing, G. S., Sampson, A. K., Kratzert, F., and Frame, J. M.: Post-Processing a Conceptual Rainfall-Runoff Model with an LSTM, EarthArXiv preprint, https://doi.org/10.31223/osf.io/53te4, 2020b.

Salas, F. R., Somos-Valenzuela, M. A., Dugger, A., Maidment, D. R., Gochis, D. J., David, C. H., Yu, W., Ding, D., Clark, E. P., and Nomanet, N.: Towards real-time continental scale streamflow simulation in continuous and discrete space, JAWRA Journal of the American Water Resources Association, 54, 7-27, 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., de Roo, A., Döll, P., Ek, M., Famiglietti, J., Gochis, D., van de Giesen, N., Houser, P., Jaffé, P. R., Kollet, S., Lehner, B., Lettenmaier, D. P., Peters-Lidard, C., Sivapalan, M., Sheffield, J., Wade, A., and Whitehead, P.: Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring earth's terrestrial water. Water Resources Research, 47 (5), https://doi.org/10.1029/2010WR010090, 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.