We would like to express our gratitude for 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 editor and reviewers on this paper.

We are pleased that both the reviewers and editor have accepted our proposed toolkit as a novel, valuable contribution to the field of hydrological modelling. Overall, we have addressed all the concerns of the reviewers and editor through this document. Here, we have included the point-by-point response to all comments and concerns. The original 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 the comments refer to the original manuscript, while the line (L) and page numbers (P) in the responses refer to the revised manuscript. In this document, all comments are numbered (e.g. E1, R1.1, R2.1). 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 comment (e.g. As per the reviewer comment 1.1).

**Editor:**

**E.1**

[Comment] I agree with both reviewers that the introduction and background material could be shortened. The sections on distributed modelling, fixed vs flexible, neural networks, etc. are all well written an interesting, but rather belong to some textbook material than to a paper. Hence, they should be shortened to cover the aspects that the current study wants to challenge or improve.

[Response] As per the suggestion, the introductory section of the revised manuscript is shortened by only keeping important content for the core of the paper. The content in the revised manuscript is arranged under five sections (Introduction, Methodology, Application of MIKA-SHA, Discussion, Conclusions).

[Changes]

- The sections on ANNs and physics informed neural networks are removed.
- No separate sections on distributed modelling or fixed vs flexible modelling are included. Only the relevant parts of those sections are moved to other sections appropriately.

**E.2**

[Comment] I think framing this study in the context of distributed modelling is unwarranted. First, the application appears to focus on simulation on a single outlet. I would argue that a distributed model, at a minimum, requires testing model predictions on multiple gauges. Second, as the authors mention in their paper, building a distributed model requires several layers of decisions. But here, the authors focus only on one of such layers, namely the model structures for each HRU.

[Response] We understand the concern raised here. The unique advantage of distributed models lies in their ability to predict runoff at multiple gauges. As with any other semi-distributed model, MIKA-SHA induced rainfall-runoff models can predict runoff at multiple gauges (at every subcatchment outlet). However, for the catchment used in the original manuscript (Red Creek catchment), we do not have internal discharge data to evaluate the internal prediction capabilities of MIKA-SHA induced models. Therefore, we have applied MIKA-SHA to a different catchment (Rappahannock River basin) where internal discharge data are available and included the results in the revised manuscript. Here, we have compared the runoff prediction accuracy of MIKA-SHA induced models both at catchment outlet and subcatchment outlets.

Even though more weightage is given for model structure identification for each HRU in MIKA-SHA, we follow the three basic steps in distributed model building as given by Fenicia et al. (2016).

Step 1: Implement a spatial discretization scheme – We identify functionally different land segments using HRUs through watershed delineation.

Step 2: Define the model structure and the connections between the spatial elements – At this stage GP-based machine learning algorithm of MIKA-SHA automatically identify the model structure for each HRU.

Step 3: Achieve model parsimony – We consider the model parsimony in terms of the number of associated model parameters as one of the evaluation criteria in the optimal model selection from the competitive models. Even though we present the model structure identified by MIKA-SHA as it is in the result section, we highly encourage innovative human cognition at this stage to improve the induced model structure.

[Changes]

- The "Application of MIKA-SHA" section is replaced with a different MIKA-SHA application to facilitate hydrograph prediction evaluation at multiple gauges (L469, P18 – L635, P28).

**E.3**

[Comment] I also think that the benefit of the proposed models would appear clearer if there was a benchmark of comparison. I agree with Reviewer 1 that this benchmark could be a lumped model at the same location.

[Response] As mentioned by you, we understand the importance of benchmarking the proposed MIKA-SHA toolkit. Hence, we compared the proposed toolkit (semi-distributed models) with two existing hydrological models (XINANJIANG and HyMOD) at the lumped setting for the same catchment. Further, we compared MIKA-SHA induced models with lumped models identified by our previously developed lumped model induction toolkit (ML-RR-MI).

[Changes] Performance comparison between MIKA-SHA induced models and lumped models is included in the revised manuscript (Section 3.2.3) (L593, P26 – L636, P28).

**Reviewer 1:**

**R1.1**

[Comment] The introduction parts are too long. Section 1.1 could be simplfied. In addition, after I read through your paper, you are using two conceptual distributed models SUPERFLEX_TOPO_M2 (Figure 5) and the FUSE_TOPO_M1 model (Figure 8). It looks like there is nothing related to ANN. If my understanding is correct, your section 1.2.1 ANN could be removed or greatly simplified as well.

[Response] As per your suggestion, we have shortened the introduction section significantly in the revised manuscript.

[Changes]

- The section on ANN is removed (including physics informed neural networks).
- Restructured the introduction section by shortening the content.

    1 Introduction (L33, P2 - L239, P8)

        1.1 Uniqueness of the place (concerning the first objective of the study) (L77, P3 – L108, P4)

        1.2 Choice of the model (concerning the second objective of the study) (L109, P4 – L128, P5)

        1.3 Machine Learning in Water Resources (L129, P5 - L239, P8)

            1.3.1 Genetic Programming (GP) (L172, P6 – L193, P7)

            1.3.2 Physics Informed Machine Learning (L194, P7 – L239, P8)

**R1.2**

[Comment] Line 170-172. "the state of art machine learning capabilities have not been tested in hydrological modeling and they expect even distributed hydrological models are to be developed primarily on machine learning in near future." Several rainfall-runoff modeling studies using deep learning have applied the semi-distributed structure using different approaches, for example:

- Neural Runoff Model from Xiang et al.: https://www.sciencedirect.com/science/article/abs/pii/S1364815220301900
- HydroNets from Google: https://ai.googleblog.com/2020/09/the-technology-behind-our-recent.html

[Response] We appreciate highlighting the recent developments in distributed rainfall-runoff modelling using DL.

[Changes] The two given research studies are cited in the revised manuscript (L167 – L168, P6).

**R1.3**

[Comment] You are using the Genetic Programming, a machine learning approach, to tune the parameters of conceptual distributed models. However, people are using many statistically-based approaches such as SUFI2, ParaSol, MCMC to calibrate the model. (https://www.sciencedirect.com/science/article/abs/pii/S0022169408002370)

So, why are we using Genetic Programming? Does it really perform better than SUFI2 or other approaches?

[Response] As you have given here, there are many statistical-based approaches for parameter tuning. We see no disadvantage of using such methods. We only provide an alternative machine learning framework that is capable of tuning parameters plus model structures. The selection of GP for the current study is associated with the primary objective of the study, which is to induce readily interpretable models through machine learning by incorporating basic hydrological knowledge and not because of it performing better (which may be or may not be the case) than statistical approaches in parameter tuning. GP has been selected as the ML technique due to its ability to generate explicit mathematical relationships among independent (forcing) and dependent variables. Therefore, incorporating hydrological knowledge can be done more explicitly with GP than other black-box type ML techniques. Further, with GP, there is no requirement for pre-definition of a model structure. Instead, identifying an appropriate model structure is part of the ML framework.

[Changes] A paragraph highlighting the rationale behind the selection of GP for the current study is included in the revised manuscript (L266 – L273, P9).

**R1.4**

[Comment] It is noticed that you have proposed ML-RR-MI and published a paper titled "Hydrologically Informed Machine Learning for Rainfall-Runoff Modeling: A Genetic Programming-Based Toolkit for Automatic Model Induction" on Water Resources Research. The difference and improvement between these two papers should be one of the key descriptions. And I have a question that did you applied the lumped model without a semi-distributed structure at the same catchment? How the performance of the semi-distributed model comparing to the model without it? The comparison would be helpful to understand the effectiveness of the semi-distributed structure in this research.

[Response] We highly appreciate highlighting the importance of comparing MIKA-SHA model performances with other approaches, such as non-machine learning and lumped modelling. As per your suggestions, we have made the following comparisons in the revised manuscript.

- Two existing hydrological models (XINANJIANG and HyMOD) are calibrated at the lumped setting for the same catchment (please note that the MIKA-SHA is applied to a different catchment in the revised manuscript to address one of the editor's suggestion) using a non-machine learning algorithm (Dynamically Dimensioned Search). Then the results are compared with MIKA-SHA induced models (lumped, fixed structure, non-machine learning vs semi-distributed, flexible, machine learning).
- Our previously introduced ML-RR-MI toolkit is applied for the same catchment to induce lumped models. Then the results are compared with MIKA-SHA induced models (lumped, flexible, machine learning vs semi-distributed, flexible, machine learning).
- Results of ML-RR-MI induced models and two existing fixed hydrological models are also compared (lumped, flexible, machine learning vs lumped, fixed structure, non-machine learning).

[Changes] Model comparison results are included as a separate section in the revised manuscript (Section 3.2.3) (L593, P26 – L636, P28).

**Reviewer 2:**

**General Comments**

**R2.1**

[Comment] The authors have provided an extensive literature review on the subject matter. However, some of the topics are less relevant and may lead astray for the reader from the main subject matter. For example, it could suffice to present the literature on machine learning applications in water resources simply in one paragraph as part of the introduction section than providing own literature review section (section 3). Similarly, the sub-section focused on lumped and distributed models can be removed from the manuscript since this is too common topic in hydrology.

[Response] As per your suggestion, we have restructured the literature review section by including only the relevant matter under the Introduction section. There is no separate literature review for machine learning applications in water resources and included under the same introduction section. The sub-section on lumped vs distributed has also been removed in the revised manuscript.

[Changes] Restructured the literature review section by shortening the content and presented it under the Introduction section.

- 1 Introduction (L33, P2 - L239, P8)
  - 1.1 Uniqueness of the place (concerning the first objective of the study) (L77, P3 – L108, P4)
  - 1.2 Choice of the model (concerning the second objective of the study) (L109, P4 – L128, P5)
  - 1.3 Machine Learning in Water Resources (L129, P5 - L239, P8)
    - 1.3.1 Genetic Programming (GP) (L172, P6 – L193, P7)
    - 1.3.2 Physics Informed Machine Learning (L194, P7 – L239, P8)

**R2.2**

[Comment] On the other hand, less coverage was given to details of certain methodologies followed in this research. For example, it would have been more helpful to provide the reader with further details on set up and components of the Genetic Programming (GP), FUSE and SUPERFLEX by removing the literature review on less relevant topics including the sub-section focused on Artificial Neural Networks (ANN) (since ANN was not used in this research). Generally, with the exception of the sub-topics focused on GP, FUSE, and SUPERFLEX the remaining contents of Section 2 (Fundamental approaches in Hydrological modelling) and Section 3 (Machine Learning in Water Resources) can be either omitted, or placed under the introduction or discussions sections in a concise and relevant form.

[Response] As per your suggestion, the section on ANN is removed. There are no separate sections under "Fundamental approaches in hydrological modelling" and "Machine learning in water resources" in the revised manuscript. Relevant contents in those sections are moved to the "Introduction" and "Discussion" sections. Then, all the methodologies used in developing the proposed toolkit is presented under the "Methodology" section. Workflow of the MIKA-SHA is explained step by step by giving more details on GP, FUSE and SUPERFLEX as you have suggested.

[Changes] The revised manuscript has the following structure (please note that the MIKA-SHA is applied to a different catchment in the revised manuscript to address one of the editor's suggestion).

- 1. Introduction (L33, P2 - L239, P8)
  - 1.1 Uniqueness of the place (L77, P3 – L108, P4)
  - 1.2 Choice of the model (L109, P4 – L128, P5)
  - 1.3 Machine Learning in Water Resources (L129, P5 - L239, P8)
- 2 Methodology (L240, P8 – L468, P17)
  - 2.1 MIKA-SHA workflow (L279, P10 – L416, P15)
  - 2.2 Purpose-built functions (L417, P15 – L456, P16)
  - 2.3 Performance Measures (L457, P16 – L468, P17)

- 3 Application of MIKA-SHA (L469, P18 – L635, P28)
    - 3.1 Study Area (L473, P18 – L507, P21)
    - 3.2 Results (L508, P21 – L635, P28)
- 4 Discussion (L636, P28 – L727, P31)
    - 4.1 MIKA-SHA_SUPERFLEX (L637, P28 – L664, P29)
    - 4.2 MIKA-SHA_FUSE (L665 – L683, P29)
    - 4.3 Model induction capability of MIKA-SHA (L684, P29 – L727, P31)
- 5 Conclusions (L728, P31 – L752, P32)

**R2.3**

[Comment] The methodology and scientific background contents appear blended in many sections of this manuscript. Thus, I would recommend having a separate Methodology section with only those methodologies followed in your study placed under this section.

[Response] As your suggestion, a separate Methodology section by including all approaches used in developing MIKA-SHA is added in the revised manuscript.

[Changes] Arrangement of the Methodology section of the revised manuscript is as follows.

- 2 Methodology (L240, P8 – L468, P17)
    - 2.1 MIKA-SHA workflow (L279, P10 – L416, P15)
        - 2.1.1 Data Preprocessing (L284 – L293, P10)
        - 2.1.2 Model Identification (L294, P11 – L328, P12)
        - 2.1.3 Model Selection (L329, P12 – L378, P14)
        - 2.1.4 Uncertainty Analysis (L379, P14, L405, P15)
    - 2.2 Purpose-built functions (L417, P15 – L456, P16)
        - 2.2.1 SUPERFLEX Library (L423 – L433, P15)
        - 2.2.2 FUSE Library (L434 – L444, P16)
    - 2.3 Performance Measures (L457, P16 – L468, P17)

**R2.4**

[Comment] Similarly, the 'Discussions' section is missing and some of the paragraphs in sections 2 to 6 appear more suited to the discussions section. Under this section, you may compare and contrast these previous research works in relation to yours with regards to the methodologies followed and results obtained in your research.

[Response] As per your recommendation, a separate Discussion section is added to the revised manuscript. The relevant content from sections 2 to 6 in the original manuscript is moved to the discussion section (please note that the MIKA-SHA is applied to a different catchment in the revised manuscript to address one of the editor's suggestion).

[Changes] The Discussion section in the revised manuscript is arranged as follows.

4 Discussion (L636, P28 – L727, P31)

4.1 MIKA-SHA_SUPERFLEX: This section discusses the model inferences made out from the optimal model identified by the MIKA-SHA SUPERFLEX library (L637, P28 – L664, P29).

4.2 MIKA-SHA_FUSE: This section discusses the model inferences gained from the optimal model identified by the MIKA-SHA FUSE library (L665 – L683, P29).

4.3 Model induction capability of MIKA-SHA: This section discusses all the general finding of the current study (L684, P29 – L727, P31).

**R2.5**

[Comment] Although the authors have effectively applied machine learning methods for model structure identification under distributed setting, I am a beat skeptical on some of the conclusions arrived in relation to the methodologies followed. For example, the available dataset was divided into four categories, i.e. spin-up, calibration, validation, and test. From the manuscript it can be noticed that both model calibration and validation datasets were used in training the hydrological and machine learning models (e.g. L448 and L464 in section 5.3). Thus, out of the total length of the dataset (i.e. 11 years), only one year was allocated for model testing (2013-2014) or for actual validation of the hydrological model. The question would then be if we can conclude that the proposed methodology achieved the intended goal. Application of the hydrological model for a single hydrologic year may provide the possibility to assess the dominant hydrologic processes in relation to the prevalent climatic and physiographic conditions in that particular year. But it would have been more helpful to use multiple single testing years, for example, using a cross-validation technique. This way the reader may get a better insight into the resulting model structures and the model test results under the different conditions and there by a more robust model evaluation.

[Response] We agree with your comment here. As you correctly said, we use both calibration and validation periods for the optimal model selection. Model training (optimizing both model structure and parameters) is carried out using only the calibration period data. Then, we evaluate the model performances of all identified models by GP-based machine learning framework on the validation period. We use this method as a way of removing overfitted models to the calibration period. Hence, the testing period efficiency values demonstrate the out of sample performance of the optimal model. So, we agree that the length of the testing period used in the original manuscript is insufficient. Hence, in the revised manuscript, the testing period has extended to four years. We are continuously working on improving the proposed MIKA-SHA toolkit.

Therefore, your suggestion on using multiple single testing years through cross-validation will be considered in the next version of our toolkit. We believe that will help to handle data scare situations well.

[Changes] Testing period is extended to four years, and 15 years of data are used in the revised manuscript (L485 – L487, P19).

**R2.6**

[Comment] Similarly, the uncertainty analysis procedure lacks information on how the parameter bounds and threshold for behavioral models are set. A threshold NSE value of 0.6 was used in this study, which I think is very low for many practical applications of a hydrological model. Capability of the prediction bounds in bracketing the observed values is inversely related with the threshold NSE value. This may have yielded to the low modelling uncertainty (high percentage of bracketed observations) of the selected model structures reported in this manuscript.

[Response] As you correctly stated, uncertainty estimation is sensitive to parameter ranges, likelihood estimator and likelihood threshold used. In the current study, we selected parameter bounds from the cited literature in the manuscript (some authors were contacted directly). Similarly, the NSE value of 0.6 is also chosen by studying previous GLUE applications. Inside our MIKA-SHA framework, we use the existing hydrological elements (in the current case, building blocks of FUSE and SUPERFLEX frameworks) as supporting materials. Our intention was not to go into details of those supporting materials or compare them. More importantly, the proposed framework can be coupled with any internally coherent collection of building blocks.

Therefore, setting parameter bounds, selecting an objective function for GLUE and setting threshold value to identify behavioural models are all user inputs of the MIKA-SHA framework. User has the total flexibility of playing with these input parameters depending on their application.

[Changes] All user inputs related to uncertainty analysis are also added to the algorithmic settings table of MIKA-SHA (Table 4) to demonstrate that the user has the freedom of choosing desired values (P21). Accordingly, the uncertainty analysis section was modified in the revised manuscript (L379, P14, L405, P15).

**R2.7**

[Comment] It would also make easier for readers who are less familiar with GP to follow the manuscript if the GP terminologies can be re-written in hydrological context. For example, what do we mean by initial population here? is it a particular hydrological model component from FUSE/SUPERFLEX ? or is it a set of hydrological model parameters ?

[Response] We understand your concern here. Hence the GP terminologies are explained in the hydrological context in the revised manuscript. Initial population means a set of candidate model structures (semi-distributed model structures made from the purpose-built functions, basic mathematical functions and random constants) that are randomly generated to capture the watershed's runoff dynamics. These model structures (GP individuals) may differ from each other in terms of model structural components and parameter values.

[Changes] Functions of GP-based machine learning algorithm of MIKA-SHA (i.e. Model Identification stage) is explained step by step by giving more details in the revised manuscript (L294, P11 – L328, P12).

**Individual Comments**

**R2.8**

[Comment] L6- 'limited use in scientific fields' seems too broad area to comment on. Consider rephrasing it, e.g. ' in rainfall runoff modelling' (accompanied by a relevant reference)

[Response] We agree that the term may be too broad. We made the above statement based on the following literature.

"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).

[Changes] The sentence is slightly modified as follows, and the above citation is added.

"Despite showing a great success of applications in many commercial fields, machine learning and data science models generally show a limited use in **many** scientific fields, including hydrology **(Karpatne et al., 2017)**." (L6 – L7, P1)

**R2.9**

[Comment] L15- rephrase 'decreasing meaningfulness of lumped models'. Lumped models might be preferable under certain conditions, e.g. for very small catchments in data scarce areas where distributed or semi-distributed model settings might be less practical.

[Response] Yes, lumped models are more appropriate for very small catchments in data-scarce situations. However, for larger catchment where spatial heterogeneity may be significant, inferences made out from a lumped model may be limited or unrealistic (still may provide good prediction accuracies).

[Changes] The sentence was rephrased as follows in the revised manuscript.

"The meaningfulness and reliability of hydrological inferences gained from lumped models may tend to deteriorate within large catchments where the spatial heterogeneity of forcing variables and watershed properties is significant." (L15 – L17, P1)

**R2.10**

[Comment] L20- 'without any subjectivity in model selection' seems less realistic since all model selection algorithms involve certain level of subjectivity, albeit at varying degrees. In your case, for example, setting the model parameter bounds (for the hydrological model) and many of the constants and assumptions related to set up of the machine learning model shown in Table 3 involve certain level of subjectivity.

[Response] Yes, we agree with your statement here. We wanted to highlight here that in contrast to traditional hydrological modelling, there is no requirement to pre-define a model structure with MIKA-SHA.

[Changes] The sentence was modified as follows in the revised manuscript.

"MIKA-SHA captures spatial variabilities and automatically induces rainfall-runoff models for the catchment of interest without any explicit user selections." (L19 – L20, P1)

**R2.11**

[Comment] L35- 'Therefore, the final goal of any successful hydrological model must be based on a physically meaningful model architecture along with a good predictive performance' But the measure of success for a hydrological model may vary from one model to another depending on the specific purpose for which they are developed. For example, physically based models might be tailored to enhance our understanding of the underlying physical system. While conceptual models might be expected to have only a partial understanding of the processes with the main purpose being to yield predictions within the required acceptable accuracy for the intended purpose. Further, blackbox models, though with little or no understanding of the underlying physical system, still have their own merits when the main goal of the modeler is just to get acceptable outputs from the set of inputs as you've mentioned in L212.

[Response] Yes, we agree with your comment here. We made the above statement based on the following literature.

"While a common end-goal of data science models is the generation of actionable models, the process of knowledge discovery in scientific domains does not end at that. Rather, it is the translation of learned patterns and relationships to interpretable theories and hypotheses that leads to advancement of scientific knowledge." – Karpatne et al. (2017)

[Changes] The sentence was modified as follows in the revised manuscript.

"**Ideally**, the final goal of any successful hydrological model must be based on a physically meaningful model architecture along with a good predictive performance." (L42 – L44, P2)

**R2.12**

[Comment] L36- 'Data science models'. Do you mean data-driven-models? Provide reference for this sentence.

[Response] Yes, we use both terms interchangeably. We made this statement based on the following reference.

"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).

[Changes] Above reference is added in the revised manuscript (L148 – L149, P5).

**R2.13**

[Comment] Section 2.3 – remove this section or take selected points from this section and concisely discuss in relation to your methodology, results or conclusions (under the Discussions section).

[Changes] Section 2.3 is removed in the revised manuscript while moving relevant content to other sections.

**R2.14**

[Comment] L167-174 – provide references.

[Changes] As mentioned above the Section 2.3 is removed in the revised manuscript.

**R2.15**

[Comment] L212- 'Certainly, if we are only interested in better forecasting results then, the machine learning models might be the preferred choice over the conceptual or process-based models due to their better predictive capability'. But can we give this generalization in light of the multiple factors affecting the relative performance of machine learning models, including length of the training dataset and nature of the training algorithm? Provide reference.

[Response] As you correctly said, the use of machine learning models may not be possible (or limited) in data-scarce situations. We made the above statement based on the following reference.

"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)

[Changes] Sentence is modified as follows, and the above citation is added in the revised manuscript.

"Certainly, if we are only interested in better forecasting results then, the machine learning models might be the preferred choice over the conceptual or physics-based models (provided no data scarcity) due to their better predictive capability (Nearing et al., 2020)." (L140 – L142, P5)

**R2.16**

[Comment] L215- 'actionable models'. Rephrase in hydrological context.

[Changes] Used the term "implementable models" in the revised manuscript.

**R2.17**

[Comment] L222- 'Further, data science models…' Do you mean: However, data… This seems to contradict with your previous statements in L218: '… offer two reasons for the limited success of data driven models'

[Response] Yes, the term "However" is the correct word. The sentence about the limited success is a general statement. Through this sentence (L222), we wanted to show the potential of machine learning models offered in hydrological modelling over the traditional models. Efforts like these may help machine learning models to achieve wide success in hydrological modelling in future.

"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 modelling." – Nearing et al. (2020)

[Changes] The term "However" is used in the revised manuscript (L152, P5).

**R2.18**

[Comment] L286-288- sentence not clear. Re-write, for example, as: Individuals with better performance (based on the objective function values) are assigned higher probability of selection and thereby given the chance to create offspring through genetic operators (crossover, mutation, and elitism).

[Changes] The sentence is re-written as follows in the revised manuscript (L183 – L185, P6).

"Following that, genetic operators including mutation and crossover are performed on current generation GP individuals to produce offspring in the next generation. The procedure for selecting parent individuals for breeding guarantees that more fit individuals have a better chance of being chosen."

**R2.19**

[Comment] L306 – mentioned?

[Changes] Corrected in the revised manuscript (L212, P7).

**R2.20**

[Comment] L306-310 – long sentence, re-write with shorter sentences.

[Changes] Long sentence is divided into two sentences in the revised manuscript as follows (L212 – L216, P7).

"Only a few explainable artificial intelligence utilizations in hydrological modelling are reported in the past (Cannon and Mckendry, 2002; Keijzer and Babovic, 2002; Fleming, 2007). However, there is an increasing trend of adopting TGDS models for recent water resources applications (McGovern et al., 2019), such as hydroclimatic model building (Snauffer et al., 2018), automated model building (Chadalawada et al., 2020) and hydrologic process simulation (Solander et al., 2019)."

**R2.21**

[Comment] L329- regularized? regular?

[Response] Yes, "regular" is the correct term.

[Changes] Corrected in the revised manuscript.

**R2.22**

[Comment] L355- 'GP has been selected as the machine learning technique here due to its ability to optimize both model configuration and model parameters together'. Was GP used to simultaneously optimize both the hydrological model structure and parameters in this study? If so, how was GP used for parameter optimization of the hydrological model? (The illustrated procedure was focused on model structure). If not, how were the hydrological model parameters optimized before conducting the uncertainty analysis (UA)?

[Response] Yes, GP was used to optimize both model structure and parameters simultaneously. Inside MIKA-SHA, elements of hydrological knowledge are incorporated as purpose-built functions (special functions), such as FUSE, SUPERFLEX and DISTRIBUTED. Functional arguments of these functions represent model structural decisions, such as the presence or absence of a particular reservoir, and model parameters, such as discharge coefficients. Therefore, when the GP individuals made out of these special functions are subjected to genetic operations (crossover and mutation), both model structure and associated parameter values are optimized together.

**R2.23**

[Comment] L377- how was the shape parameter value (2.5) of the Gamma-distribution based routing function determined?

[Response] Shape parameter is fixed at 3 (not 2.5) as did in the original FUSE paper (Clark et al., 2008).

[Changes] Corrected shape parameter value to 3, and the citation mentioned above is added in the revised manuscript (L453 – L454, P16).

**R2.24**

[Comment] L425- Elaborate on how the number of independent runs and other algorithm settings of your framework (Table 3) were determined. For example, why not 10 or 30 independent number of runs instead of 20? Similarly, why a generation number of 50 or population size of 2000 etc.

[Response] In general, there are no specific rules to select the appropriate GP algorithmic settings. Usually, settings are chosen through a trial and error procedure (prior experience can also be used here). By being an evolutionary algorithm, higher population sizes, higher generations, and higher independent runs may result in more accurate models due to higher diversity and model evaluations. However, the chosen settings eventually decide the computational time and demand. Hence, one can select GP algorithmic settings based on the available computing power and time.

[Changes] Above explanation is added in the revised manuscript (L287 – L289, P10).

**R2.25**

[Comment] L471 - The selection of Cross-sample entropy parameters (e.g. r) are quite critical for the evaluation result. How were these values determined in your study?

[Response] Again, we have taken Cross-sample entropy parameters from the cited literature in the manuscript. As you correctly stated, these parameter values are critical for the result evaluation. However, inside the MIKA-SHA framework, Cross-sample entropy is used as a relative performance measure to evaluate time-series complexities among competitive models. Therefore, once the parameters are set, they are common to every competitive model.

**R2.26**

[Comment] L487- what are these model parameters? elaborate and provide reference.

[Response] These are the model parameters of MIKA-SHA-induced semi-distributed models (either SUPERFLEX or FUSE). Currently, model parameter details are given in the Appendix.

[Changes] References for FUSE and SUPERFLEX frameworks' parameters are given in the revised manuscript (L755, P32 & L759, P34).

**R2.27**

[Comment] L498- '…is changed uniformly…'. Do you mean: …are generated … ?

[Response] Inside MIKA-SHA, all parameter values are normalized between 0 and 1. Therefore "changed uniformly" means assigning a value between 0 and 1 from a uniform distribution.

**R2.28**

[Comment] L500- how many and which model parameters were allowed to vary and how were these parameters selected out of the total number of model parameters?

[Response] Number of parameters changed and selection of parameters to change at a time are randomly decided. Hence, at every iteration different number of parameters are changed. For example, if there are 20 model parameters, first, a number between 1 and 20 is randomly generated from a uniform distribution (say the generated number = 10). Then, ten integer numbers between 1 and 20 are randomly generated from a uniform distribution. Parameters corresponding to those numbers are subjected to change at this iteration. This process is repeated until a user-defined number of behavioural models are identified.

[Changes] An explanation is added in the revised manuscript (L388 – L389, P14).

**R2.29**

[Comment] L500-'…while keeping the remaining model parameters at their calibrated values.'. How were these model parameters calibrated before conducting the UA. Was GP applied for that purpose? It would be helpful for the reader if you can clarify this in relation to the comment mentioned under L355.

[Response] As mentioned in R2.22, GP is used to optimize both model parameters and structure simultaneously.

[Changes] A sentence about how the model parameters are calibrated is included in the revised manuscript (L384 – L385, P14).

**R2.30**

[Comment] L501- why was this threshold NSE value of 0.6 chosen?

[Response] We selected an NSE value of 0.6 from previous GLUE applications like Beven and Free (2001). However, the likelihood threshold and likelihood estimator (objective function) are user inputs of the MIKA-SHA framework.

[Changes] Citation mentioned above is added in the revised manuscript (Table 4, P21).

**R2.31**

[Comment] L502- The term 'behavioral models' in this case refers to the parameter sets rather than to their NSE or discharge values.

[Changes] Corrected in the revised manuscript (L387, P14).

**R2.32**

[Comment] L512- '…to measure the uncertainty estimation capability of the selected optimal model'. Do you mean: … to measure the level of modelling uncertainty of the selected optimal model structure? In your study, the GLUE methodology was used to estimate the level of uncertainty, while the selected optimal model structure was itself subjected to uncertainty analysis rather than being used as an uncertainty estimation tool.

[Response] Uncertainties of model predictions are unavoidable in any hydrological modelling exercise due to the lack of understanding of the underlying physical phenomena and the inability to measure accurately (Beven, 2012). Measurement errors, errors due to model simplifications, inaccuracy of process representation, parameter uncertainty are the sources of modelling uncertainty. In this approach, we use GLUE to identify the percentage of bracketed observations by changing model parameters while keeping a satisfactory model performance. Hence, we use the percentage of observations that lie between the 90% uncertainty bands to check whether the parameter uncertainty of the optimal model alone can account for total model uncertainty.

**R2.33**

[Comment] L513- 'If the uncertainty estimation capabilities are satisfactory, the model performance of the optimal model is tested for an independent time frame (2013/01/01 to 2014/12/31) which is not used in model selection or identification stages'. Re-write this sentence as well in accordance to the previous comment. What was your criteria for a satisfactory level of modelling uncertainty (or as in your text- a satisfactory uncertain estimation capability)? It seems that both the model selection and validation periods were actually used for model identification (selection) and not for a hydrological model validation or testing. And a single year of model testing looks quite short period to arrive at a conclusion. Thus, if you have data limitation from additional periods or if some of the hydrological model identification (selection) years cannot be moved to the hydrological model testing (validation), you may consider using alternative model evaluation techniques such as the leave-one-out or other cross-validation techniques. This way you may get more validation (test) results that can help you arrive at a relatively robust conclusion.

[Response] As mentioned above, we use GLUE to identify the percentage of bracketed observations by changing model parameters while keeping a satisfactory model performance. Hence, we use the percentage of observations that lie between the 90% uncertainty bands to check whether the parameter uncertainty of the optimal model alone can account for total model uncertainty. Again, the satisfactory level is a user-defined value. A bracketed observation percentage above 60% was identified as a satisfactory uncertain estimation capability in this study.

As you correctly said, we use both calibration and validation periods for the optimal model selection. Model training (optimizing both model structure and parameters) is carried out using only the calibration period. Then, we evaluate the model performances of all identified models by GP-based machine learning framework on the validation period. We use this method as a way of removing overfitted models to the calibration period. Hence, the testing period efficiency values demonstrate the out of sample performance of the optimal model. So, we agree that the length of the testing period used in the original manuscript is insufficient. Hence, in the revised manuscript, the testing period has extended to four years. We are continuously working on developing the proposed toolkit. Therefore, your suggestion on using multiple single testing years through cross-validation will be considered in the next version of our toolkit. We believe that will help to handle data scare situations well.

[Changes]

- An explanation about what we mean by satisfactory uncertain estimation capability is added in the revised manuscript (L397 – L398, P14).
- The testing period is extended to four years (L485 – L487, P19).

**R2.34**

[Comment] L626-629 – 'Out of the 33 model parameters only 5 parameters can be identified as sensitive parameters. … This demonstrates a lesser dependency on model parameters compared to the total model performance in semi-distributed modelling owing to the large number of model parameters.'. Among other factors, sensitivity analysis results depend on the minimum and maximum values of a parameter dimension. How, were these values fixed in your study?

[Response] Yes, as you correctly said, parameter ranges affect the sensitivity analysis results. In the current study, we used the parameter values from the cited literature in the manuscript. However, the user has the flexibility to change them accordingly.

**R2.35**

[Comment] L629-'FUSE_TOPO_M1 results in high value (94%) for the percentage of measured streamflow data within the confidence interval bands and hence shows a significant capability of estimating associated uncertainty.' Re-write this sentence in accordance to the comment provided in L512. The percentage of observation bracketed by the uncertainty bounds is highly dependent on the threshold value used during behavioral model identification. The threshold NSE used in this study (0.6) seems very low as compared to the reported calibration and validation results of the optimal model. Given this low threshold NSE, it is expected to get a high percentage of the observations falling within the uncertainty bounds. Thus, try to justify why you adopted this threshold NSE value (under the methodology section).

[Response] As mentioned above, we picked the NSE value = 0.6 based on literature (Beven and Free, 2001). Yes, we agree that the NSE value = 0.6 as the likelihood threshold is low compared to MIKA-SHA-induced models' performance for the Red Creek catchment. In the revised manuscript, we have applied MIKA-SHA to a different catchment, and as per the MIKA-SHA induced models' performance for the new catchment, we believe NSE value = 0.6 is appropriate.

[Changes] Above citation is included in the revised manuscript (Table 4, P21).

## REFERENCES

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

Beven, K., and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, Journal of hydrology, 249(1-4), 11-29, 2001.

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.

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.

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, 2020.

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.