Reply to RC1: 'Comment on egusphere-2025-1699', Georgios Blougouras & Shijie Jiang (co-review team)

We sincerely would like to thank the review team for their detailed and constructive evaluation of our paper. In this document we reply to their comments.

General Comments

The manuscript of Álvarez Chaves et al. tackles a highly relevant, important and immediate issue that the hybrid hydrological community is facing. The authors explore and compare the role of the data-driven component across different hybrid models, using entropy-based metrics, in two experiments: a synthetic experiment (known ground truth) and a real-world experiment (provided by the CAMELS-GB dataset). Does the conceptual model, to which the data-driven component is attached to, provide any guidance? Or is the data-driven component overpowering it (potentially 'overriding' parts of the conceptual model along the way)?

I found the article very informative. As I mentioned before, the question at hand is highly relevant, and the authors make a great effort in developing a robust methodology to explore it. The methodological contribution is, in my opinion, not limited to the proposed metric, but also extends towards the authors' diligence to explore everything that happens 'under the hood' of their models. I expect the findings to resonate with hybrid hydrological modellers. At the same time, the suggested methodological workflow is of interest to more process-oriented/catchment-scale hydrologists as well, since it offers a way to test and refine the physical representation of their models.

The 'personal' language that the authors use (comments in parentheses, quotes inspired by general literature etc.) is also appreciated. It is dynamic and engaging, which enhances the reading experience. The manuscript is, in general, well formatted and easy to follow (with some exceptions - refer to 'specific comments'), and has clear figures.

Thank you for your nice comments! We really appreciate your careful reading of the paper and we are very pleased that you found it engaging.

However, there are some key conceptual/general understanding concerns (see next paragraphs of this section), as well as some more specific comments (see the relevant section below) that should be addressed by the authors, before the manuscript is ready for publication.

General comment 1: According to the phrasing of the paper in many instances, the readers might be led to believe that hybrid modeling is used here as a way to help the LSTM predictive performance (e.g., L106-107: 'While purely data-driven… genuinely enhances model performance', or, L621: '...reduce the effort required…'). However, the LSTM's role in the hybrid architectures explored in the manuscript is not to 'lead the predictions', but rather to infer the conceptual model parameters. Therefore, describing the models as 'physics-constrained' (L9) might not be the most accurate description - maybe something like ML-enhanced/parameter

learning/… differentiable modeling would be more fitting. The change in terminology (although potentially annoying), implies different perspectives regarding the role of the data-driven and the physics-driven components of the model. To be more precise, the manuscript explores if different architectures are 'helping the LSTM', even though to begin with, the concept of parameter learning usually reflects the opposite direction (i.e., exploiting the LSTM to help the conceptual model improve its predictions). To my opinion, this creates a 'conceptual mismatch' across the perspective of the manuscript and the 'typical' way such hybrid models are used. I encourage the authors to clarify if their framing is applicable or not under this context - maybe it might be more accurate to view the LSTM as being 'constrained', rather than 'assisted' by the conceptual model. In light of this architecture interpretation, it would be helpful for the authors to revisit their interpretation of the entropy-based results, to ensure that the evaluation and subsequent conclusions are aligned with the actual model structure and what it represents.

Thank you for this thoughtful comment about terminology and framing.

Indeed, we show two different modeling setups: Section 4.3 shows the results of parameter learning to enhance a conceptual model as the reviewer describes it, and Section 4.4 shows the results for post-processing a regular conceptual model to improve performance of a data-driven model. As these are two very different approaches and there exists no rigorous definition of what a "hybrid model" is, we see this name as an apt description of what we did in both cases and we would like to keep it as such.

Both perspectives, led by the data-driven component and physics-constrained, are valid and depend on the researcher's background and starting point. A physics-based modeler would naturally view our Hybrid SHM as a conceptual model with dynamic parameters, while a data-driven modeler would see it as a constrained LSTM. Further, our analysis results indeed suggests rethinking the terminology: if one starts from the more traditional hydrological viewpoint of improving a conceptual hydrological model by letting an LSTM control its parameter values, the performance results with the pure LSTM scoring best in most cases will cause the modeler to frame the question rather from the LSTM side: is the additional conceptual part in any way helpful (i.e. constraining in the right way, or assisting), or rather making the prediction problem harder? In a revised version of the paper, we will carefully check the order in which we introduce and use the different terminology, and make sure that it is in proper context with our premises and results.

For the hybrid models which focus on parameter learning we do see the role of the LSTM as leading predictions. As an example: because Hybrid Nonsense modifies 'su' and 'si' to behave as time lags for the output, the outflow is mostly managed by the interaction between 'sb' and 'kb'. Considering the previous point, in Figure 8 we see that for the high flow peak that happens in 2005-01, 'kb' is increased disproportionately just so that the model can match the peak based on the volume available in 'sb'.

The role of the LSTM as leading predictions is more clear for the post-processing models, as the predictions are given directly by the data-based component being informed by the results of

an initial run of the physics-based component. Therefore we do see the two setups as cases in which the conceptual model assists (or doesn't) the LSTM. Moreover, particularly the setup that uses parameter learning can be effectively described as physics-constrained because conservation of mass is imposed by the conceptual model at the end of the pipeline. Therefore we consider that both terms: 'constrained' and 'assisted', apply.

We will add more detailed commentary on the 'sb' and 'kb' behavior in Figure 8 to better illustrate how the LSTM leads predictions, but we believe our current terminology accurately reflects the model architectures and their roles.

General comment 2: The manuscript defines the value of hybrid modeling primarily in terms of streamflow prediction, measuring the "utility" of physics constraints by the degree to which they reduce LSTM parameter variability (entropy) for this single task. In my view, this approach risks narrowing the broader motivation for hybrid models, which is not limited to improving runoff prediction or reducing model complexity, but also includes enabling more meaningful and diagnostically useful process representations (e.g., for ET, soil moisture, or system anomalies). I am concerned that by focusing only on this narrow predictive context and a single diagnostic (entropy), the study may misrepresent the broader role of physical constraints in hybrid models. Many would consider physical constraints valuable even when predictive skill does not improve, as they support interpretability and process fidelity. I recommend that the authors more explicitly clarify the intended scope of their analysis and discuss the limitations of generalizing these findings to other goals of hybrid modeling.

Our focus on streamflow prediction reflects what we observe as the predominant approach in current literature. As discussed in Sections 1.4 and 4.1, many existing hybrid modeling studies emphasize predictive performance for a single target, often without detailed analysis of why specific physical constraints or conceptual components are effective relative to alternatives.

Rather than dismissing the broader value of hybrid models for interpretability and process fidelity, our study aims to provide a methodological framework for systematically evaluating the degree to which physical constraints actually constrain model behavior. This is highly relevant also in the context of interpretability and process fidelity: if the intended model structure that shall guarantee interpretability is overwritten, this argument is no longer valid or we might actually find a better (in our case: simpler) process representation that again allows for interpretability and scientific learning. While we focus on entropy as our primary metric, we believe entropy complements, rather than replaces, other evaluation approaches focused on process representation and interpretability, as we demonstrate in Section 4.3.2. We appreciate the perspective expressed by the reviewer in this comment and will include these considerations as an introductory point for this section in a revised version of the manuscript.

General comment 3: In the current setup, the LSTM is only used to infer parameters, and streamflow is the sole observation timeseries used for evaluation. In my view, this setup may further amplify the well-known problem of equifinality in hydrological modeling. From this

perspective, I am not convinced that low LSTM entropy necessarily reflects a physically meaningful or faithful conceptual model; rather, it may simply mean that the LSTM has found a convenient way to “satisfy” the runoff constraint, possibly by exploiting compensatory effects among parameters or model components. While the manuscript explores this issue by visualizing the distributions of the LSTM-predicted parameters (e.g., in Fig2 and Fig9), I would encourage the authors to discuss more explicitly how equifinality in particular might influence the interpretation of their results.

Thank you for this insightful comment about equifinality. We do explore this phenomenon to some degree in the synthetic case study which features overparameterized models, and there we found that the LSTM identifies rather robust and simpler solutions. Yet, generally, constraining a multi-parameter model by a single objective inevitably comes with the risk of equifinality. This issue is well-known in hydrology and it has been for a long time, therefore our study setup is consistent and comparable with existing practice. We agree that moving on to higher-variate constraints is desirable, yet we prefer not commenting on this in this manuscript because we are currently exploring this connection in a separate study, as the theoretical and methodological implications extend beyond the scope of the current manuscript.

Specific Comments

Introduction: Why are you mentioning the modular hydrological frameworks more than once (L31, L47,...)? To my understanding, there is no clear and direct follow-up regarding such efforts in the rest of the manuscript (though one could make a conceptual link, but it would still benefit from a clarification from the authors).

If we see SHM as a proxy for a typical conceptual hydrological model, Bucket and Nonsense arise as alternatives that come from modular hydrological frameworks and the value of multiple working hypotheses even in a setting in which they are combined with neural networks.
In a new version of the manuscript, we will reference these ideas again in Section 2.2.2 where the hybrid models of the study are introduced. Also, inspired by this comment, we plan to place our findings in the broader context of modular hydrological frameworks in the conclusion (what can we learn from our results with respect to building conceptual models).

For the ‘1.3 Hybrid models’ subsection: To someone unfamiliar with such modelling efforts, I think the current subsection lacks a bit of clarity - why should people care about hybrid models, and how have people employed them in the (recent) past? I think a little bit more context is required in this section (especially given that in the end, from your findings and conclusions, the paper would appeal not only to ‘hybrid modellers’, but also to the general hydrological modeling community). In general, every ‘modeling introduction’ subsection (i.e., 1.1, 1.2, 1.3) should have clear pros and cons of using each modeling type - this is not clear for section 1.3. Furthermore, additional context would be beneficial regarding the hydrological community’s evolution from 1.1 to 1.2 and then 1.3. Why did LSTM become the ‘benchmark’ for many researchers in the last 5 years? Why do more and more hydrologist attempt to use hybrid models instead of pure

data-driven models? Some of these questions are already answered in the text, but in a way that in my opinion might not be extremely clear to someone that has is not well familiarized with hybrid modeling.

We believe the points you've raised are addressed within the current text: motivations for hybrid models appear in lines 85-87 and 97-99, the rise of LSTMs as benchmarks is discussed in lines 56-60, and the shift toward hybrid approaches is covered in lines 71-74.

Moreover, this study builds directly on existing hybrid modeling literature and serves as a follow-up to Acuña Espinoza et al. (2024). We designed it for readers already familiar with hybrid modeling approaches rather than as an introductory text for newcomers. Given this context and scope, we believe the current level of background information is appropriate for our intended audience.

Yet, we acknowledge your point and, in line with our previous answers, in a new version of the manuscript we will broaden Section 1.3 to highlight additional benefits of hybrid modeling beyond performance improvements, which may help contextualize our work for a wider readership.

L106: Do you mean ‘data-driven components’? Because the hydrological models inherently are ‘physics’-driven models. Otherwise, please revise.

The statement should read: “They demonstrate that incorporating physics-based components or prior knowledge doesn’t yield an improvement in model performance *over the data-driven solution*”. We will modify it in a new version of the manuscript.

L105-109: This whole text passage has the point of view of someone who wants to improve data-driven hydrological models by incorporating physical principles. At the same time, the opposite pathway (trying to improve physical models by using data-driven modules) is also common in hydrology, and can also benefit from your suggested contributions. Why not mention it here as well?

Thank you for the suggestion, indeed our proposed method can go both ways. We will add this statement in a new version of the manuscript.

L110-111 (contribution 1): This quantitative metric is not ‘self-standing’, to my understanding, but relevant to a ‘benchmark’ model (in this case, the pure LSTM), right?. I believe it would benefit the clarity of the manuscript if the authors were explicit about this here

Indeed, the metric requires a benchmark data-driven model to serve as a baseline. We will add “in comparison to a purely data-driven benchmark” at the end of this sentence in a revised version of the manuscript.

L114-115 (contribution 3): This suggested contribution is somewhat unclear to understand under the provided context - a reader would need to read the following manuscript first to fully grasp what the authors mean by ‘effective’ and ‘prescribed’. I suggest revising this.

Thank you for the suggestion. To clarify, we will add “based on the conceptual model” at the end of this sentence in a revised version of the manuscript. This is what we refer to as the prescribed structure.

L122: ‘favoring’ -> again, the context is missing at this point for the readers to fully understand what ‘favoring’ the data-driven component means, and I would suggest revising this a little bit to provide more information.

Thank you for the suggestion. We will change this sentence to “High entropy points to an imbalance in which the data-driven component compensates for inadequacies in the conceptual model by manipulating its parameters...” in a revised version of the manuscript.

L173: Similar to my point regarding the 1.3 section, I think the authors should provide a little bit more context about why this revival of dynamic parameters has happened in hydrological modeling. I would refer more to Tsai et al. here (https://doi.org/10.1038/s41467-021-26107-z), which is already cited in your paper regardless.

Thank you for the suggestion. Yes, based on this and the previous comment, we will extend Section 1.3.

L174 / Figure 1: To my understanding, the different model setups are not yet utilized, and they do not become fully relevant until section 3 (which could be viewed as a gray zone between a ‘methods’ and a ‘results’ section). It is fine if the model setups remain in Figure 1 and this section, but then the authors should also present what the individual models represent, because there are 5 subfigures in figure 1, leaving a lot of questions to the readers (especially the ones not familiar with past group efforts utilizing the SHM and Nonsense models). Otherwise, the authors could immediately move the figure and refer to it in a more relevant section.

Thank you for your suggestion. We will add a brief description of each model here as we believe this is a good place for Figure 1.

NSE* calculation: In the original work, this metric is cross-basin and there was an additional sum in the formula (over all basins, which does not exist here), while Kratzert et al. divided by the number of basins instead of the number of days. Here you use a basin-average metric, deviating from the original formula as far as I can see (but please correct me if I am mistaken). Does this imply that the evaluation is done on a basin-scale? It would confuse/conflict with the rest of your paper where you indicate that you train on multiple basins.

Indeed there is a difference and we will clarify in a revised version of the manuscript.

In the way we wrote Equation 4, we are not including the terms related to batch averaging. In our training pipeline, NSE* is calculated per training batch where each batch contains training samples from a single basin letting us also calculate s_i per batch without having to consider standard deviations from more than one basin. Training on multiple basins happens at the end of a training epoch where we average the loss function for all batches. This is different from how the expression is shown in Kratzert et al. (2019b) but both loss functions are functionally equivalent.

Section 2.3: I am concerned that the entropy metric may depend strongly on the specific architecture and hyperparameters of the LSTM. Have you tested how robust the entropy-based findings are to changes in LSTM design (e.g., number of hidden units) or to using alternative machine learning models? I am curious to what extent the conclusions hold beyond the particular ML setup chosen in this study.

It is correct that the absolute values of entropy depend on the chosen ML architecture; however, the qualitative ranking between the different hybrid models remains stable. We confirmed this by previous analyses where we initially applied our method to four models using five hidden states and obtained the same qualitative ranking. When we extended the analysis to eight models, we increased the hidden nodes to ten to achieve comparable performance across all models. This is crucial because, in practice, a researcher typically selects an LSTM architecture to optimize *performance* on their chosen loss function. Since our entropy analysis is applied post hoc to trained models, the key requirement is that all models use the same architecture, rather than the specific architectural choices themselves.

L200: Here I would suggest a small change in the phrasing - maybe emphasize that you move away from analyzing the entropy metric of the predicted parameter values, but nevertheless exploring (and visualizing) the LSTM predicted parameters is an important step in exploring the 'under-the-hood' performance of the model. Currently it reads as if you will completely ignore the information from the LSTM-predicted parameters.

Good point and thank you for your suggestion! We will change the phrasing in a revised version of the manuscript.

L214: Mention that the link for the UNITE toolbox is found in the 'code availability' section.

We embedded a link to the text, but you make a better suggestion. Thank you! We will change it in a revised version of the manuscript.

L233 -> Appendix A1: please specify how these parameter ranges can be justified (you mention Beck et al., 2016, 2020 but later on in the text, leaving an open question for now).

There is actually a mistake in how the choice of parameters is described. The parameters from Beck et al. (2016, 2020) are for the case study in CAMELS-GB. From this reference, we

adapted the ranges to fit the simpler models in the didactic examples. We will rectify this and clarify our choices in a revised version of the manuscript.

Section 3.2: I think it would be beneficial to be more clear as to why you use a stand-alone LSTM model to compare against the hybrid setups (for example, L727-L730 of Appendix B could be mentioned here).

Thank you for your suggestion! We will move the description of Model 0 to the end of Section 3.1 as we're also advocating for an initial purely data-driven reference model.

L274: I would say that randomly selecting 5 basins and not repeating the experiment might not convince many large-sample hydrologists - especially if you do not elaborate on these random catchments or their representativeness… One could repeat the exact same experiment by using 5 other randomly selected basins (no need to go too much into detail, providing the results in the supplementary and the logs on the repository is more than enough), or further demonstrate why conducting an experiment on a set of 5 randomly selected catchments is more than enough to derive safe and transferable conclusions.

A key aspect of our didactic example is that we use synthetically generated streamflow time series from a conceptual model with known static parameters, rather than observed streamflow data from real basins. In this synthetic framework, the specific characteristics or representativeness of the selected basins are less critical because the "observed" streamflow is generated using controlled parameters rather than reflecting actual basin behavior. As such, we could have sourced the rainfall data from anywhere or even used a stochastic process to generate it for this example. This is noted in L279 but we will stress this point in a revised version of the manuscript.

L362: I think mentioning results about models appearing in the supplementary (and especially figures that have additional results from models not yet revealed before the supplementary information) can lead to unnecessary complication. Maybe you can either remove them from the main figures and reveal them directly in the supplementary information. Alternatively, if you wish to keep them, maybe some additional information about models 6,7, and 8 in the main manuscript would be helpful. Furthermore, you could also have specific shapes for Fig. 3 and 4 representing each type of 'wrong model type' - this would be practical if you wish to keep models 6, 7 and 8 in the main figures (e.g., over-parameterized models have a 'square' shape or a certain color, 'wrong architecture' models have striped color, etc…). This is not necessary to be done, just a suggestion / visualization experiment.

We appreciate the suggestion, however we would like to keep it as it is right now. We provide a description of how these models resemble those in the main text in L361-363, and we decided not to provide a longer description because of their similarity to the models described in more detail previously in this section. The suggestion for the change in visualization is good, but there is an overlap as overparametrized models can also have wrong processes or architecture. These categories are merely descriptive and we don't think it's necessary to modify the axis.

L380: ‘matches the true system’ (and many other instances in section 3 where the word ‘true’ is used) -> I feel like the phrasing is a bit misleading, as true is currently an idealized setup. Of course you have mentioned this in the manuscript, but still the word choice is important, and someone could make the implicit connection that these models can well capture the real word ‘true’ signal, which is not the case as we see in section 4.

Thank you for the suggestion. The first instance of the use of the word *true* to describe the idealized setup in Section 3 happens in L235 and we use quotations to emphasize that this is a case of synthetic “truth”. We reinforce this distinction in L498 for the real world case study. We believe this establishes context and a reader will keep this in mind while reading this section. While we appreciate the concern about terminology, as this issue is more philosophical, we prefer to maintain our current usage of "true" for the synthetic reference as it better serves the pedagogical aspect of this section.

L383: clarify which LSTM based entropy you refer to here - hidden-state or parameter based, because the distinction is important and it needs to be clear to the readers.

Thank you, we will add “hidden states” in a revised version of the manuscript.

‘On the Complexity of the Prediction Task’ -> I think this can be seamlessly merged with the 3.4 subsection, or go directly into the appendix - I feel like it doesn’t provide enough information as a stand-alone part of the manuscript…

We would prefer to keep this section as it reflects our interest in data-driven baselines and the analogy with complexity will appeal to some readers. Moreover, we highlight that our approach focuses on the data-driven component and measuring the entropy of the conceptual head layer is an open challenge that we hope can be addressed in the future.

Subsection 3.4 ‘Summary of the proposed approach’ and overall point about Section 3: Now all the models had a (almost) perfect fit. This makes the evaluation ‘easier’, because now we know if the LSTM had to work ‘overtime’ to ‘save’ the model performance. But how can we evaluate this in cases where the models do not have equally comparable performance metrics? In other words, how do we measure the ‘effort’ by the LSTM to 'save' the streamflow prediction, if the final models do not predict streamflow prediction equally well? I am aware you touch this a little bit on the next chapter, but I think additional discussion on this matter on Subsection 3.4 would be very helpful - I am sure many readers would have this question.

We agree that this is a practical consideration that readers will share and, in a revised version of the manuscript, we will add additional commentary to point towards the relevant section in this subsection.

However we would like to emphasize that this is a didactic example which is predicated on having equal performance because we’re not making any specific statements about model

selection or preference. Specifically we are using this section to apply our proposed method in a controlled setting, and derive insights about the amount of effort prediction takes without considering any differences in performance.

In practice, we do consider that any kind of ultimate selection should be based not only on performance and entropy, but also on other criteria such as explainability, computational cost, etc. This is why we highlight very specific cases in Section 4.3.

L454-459: Here, you could mention that a description of these models (SHM, Bucket and Nonsense) can also be found later in this manuscript.

Thank you! In a revised version of the manuscript we will reference Figure 1 and the relevant subsection.

L476: '...improving prediction skill' -> I assume you mean compared to your LSTM-baseline, right? Be more specific as the baseline comparison is important.

Indeed! We will add "in regard to the LSTM baseline" in a revised version of the manuscript.

L485: 'These five basins were carefully chosen…' -> how? Please elaborate. Also, are these 5 basins the one in Fig. 6? This is not immediately clear. And follow up question on L519-520: Could this also be related to the hydrological processes involved in these catchments? You do not provide any information regarding the catchments, despite being carefully chosen (L485). Not all models can fit all catchments, and in the manuscript it is not explained how these catchments were selected/what are their characteristics and so on… This creates an uncertainty to the readers: how do we know that these results are not affected by the catchment selection and specific catchment behavior? In other words, can the authors ensure that their findings are transferable across and beyond the 5 selected catchments? (similar question - point for selecting the basin 73014 in L568 and Fig. 9).

We agree that this is not clear and will improve it in a revised version of the manuscript. The five examples were carefully chosen basins in which all hybrid models have performance on par with the LSTM, as in our didactic example. Moreover in Fig. 8 and Fig. 9 we use basin 73014 because it belongs to this subset.

Our study focuses on demonstrating what our proposed entropy metric can reveal rather than selecting optimal models; hence, it is not relevant what exactly are the characteristics in those catchments, but rather we see them as useful examples to demonstrate what the results of our proposed analysis may look like.

L564-L567: You mention the two different components. The first being, the low vs high entropy, and the second being the unaffected vs suspicious time-varying patterns. Do you think there could be a visual representation of this on a 2-axes plot? I am thinking something like figure 4

but 2-d. In this case, the more you move towards the 'high' entropy values, the less important the change in parameters is (you already have mentioned that high entropy indicates 'struggling due to the imposed constraint'), but the more you move towards the 'low' entropy values, you can get different 'color' (if we draw a comparison to fig. 4), depending on whether we have high or low variability in the parameter axis. This is not necessary, but I thought it might be interesting to try and visualize this important insight - it could help promote and establish the detailed methodology.

Thank you for the interesting suggestion. The problem is that, as we saw in Section 3 and Figure 3, comparing the entropy in the parameters of models with different numbers of them will result in disingenuous conclusions. There are ways in which this could be avoided, but we find our methodology most simple by focusing just on the entropy of the data-driven component.

L573-574: Wouldn't the fact that the model has learned behaviors from training on other basins a good thing? I am confused about the point you aim to make here.

This is clarified in the next sentence. For this basin, SHM by itself already proved to be a good model and there is really no benefit from the hybrid approach and training in multiple basins.

L585-589 and Figure 10: You mention this later on, but it would be nice to elaborate already a bit on why the Hybrid Nonsense is the top 2 - it is a logical question that a lot of readers would immediately have.

Agreed! In a revised version of the manuscript we will start to comment here.

Fig 11: Would be helpful to add the % in the bars as well.

The relevant percentages are already used in the text and the bars already give a sense of proportion. Moreover, we used the same scale for the x-axis in Figure 11 and Figure 14, so we would argue that there is no need to include percentages in the figure.

General question about section 4: What about a joint analysis of ΔH and ΔNSE? I mentioned this earlier as well, but when trying to simulate real world catchments, the performance of the models can vary quite a lot - then it would be hard to judge and compare the different entropies across models, if the baseline of their predictive performance is not comparable. It would be nice to provide some clarifying perspectives on this while concluding section 4.

Thank you for this important point and, as we mentioned in our previous reply, a decision about model selection should account for performance and not only our proposed metric, which is tailored to diagnostics, not model selection in specific. In fact, with this manuscript we don't make any specific statements about model selection because we believe that this is an even broader topic in which not only performance, and now entropy, should be considered, but also interpretability and computational cost, for example. We will, however, comment on possible

ways to include ΔH and ΔNSE in a joint metric for model selection in a revised version of the manuscript, as “nuclei” for future research.

L648: I would like to suggest adding the word ‘evaluating’: ‘building **and evaluating** hybrid models’. It is a bit ‘nit-picky’ as a comment, but I think it is an important part of your implications and deserves to be mentioned.

Thank you for this suggestion. We also see how the word ‘evaluating’ fits here and will add it in a revised version of the manuscript.

L672: This sentence reads like you imply something along the lines of: ‘process-based modeling of catchment scale streamflow is unnecessary - why go in the long effort of creating or applying these models if LSTMs can be better?’. I know this is not your initial intent, so I would suggest rephrasing in order to avoid confusion.

Indeed, that is not our intention. We will place this better into context in a revised version of the manuscript.

Appendix B: It would be helpful to add some more context/information about the design of the additional models.

We believe that there is a clear motivation and explanation for each model, but in a revised version of the manuscript will acknowledge that they were inspired by the idea of having multiple working hypotheses and flexible modeling frameworks.

Technical Corrections

L38: ‘Typically catchment scale processes of in a rainfall-runoff…’

L174: ‘...used [in] our case…

L210: us -> is?

L211: you repeat ‘of’

L253: I believe ‘setup’ is just a noun - ‘set [space] up’ should be the verb needed in this context. Maybe I am wrong.

L351: ‘fix’ -> ‘adjust’ might be better fitting?

L408: is there a full stop [.] missing after ‘model’?

L472-475: This period is 4 lines long and quite hard to read through - I would suggest splitting up to individual sentences to ensure readability.

General ‘correction’: make sure you adopt a unified style when it comes to capitalizing titles throughout the manuscript. I see some passages have a ‘capitalized’ style (e.g., ‘Comparing Conceptual Constraints on the Entropy Axis’) and some others are more free with capitalizing words (e.g., ‘3.3.2 Measuring entropy of conceptual model parameter space’).

Thank you for your careful reading of our manuscript! In a revised version we will fix all of these technical corrections.

References

- Acuña Espinoza, E., Loritz, R., Álvarez Chaves, M., Bäuerle, N., & Ehret, U. (2024). To bucket or not to bucket? Analyzing the performance and interpretability of hybrid hydrological models with dynamic parameterization. *Hydrology and Earth System Sciences, 28*(12), 2705–2719. https://doi.org/10.5194/hess-28-2705-2024