Author’s response

Dear Dr. Fenicia,

Thank you for handling our manuscript and enabling our discussion with the reviewers.

To your specific points:

Reviewer 2 acknowledges improvements but finds critical logical issues remain, warranting rejection in the current form. There is a concern that entropy is oversimplified: low entropy does not guarantee physical adequacy nor high entropy inadequacy. The synthetic case needs added constraints, dynamic vs. static parameter handling requires clarification, and claims regarding hybrid models and physics overwriting are considered misleading.

We thank the editor for their assessment. We have considered Reviewer #2's concerns and responded systematically to each point as detailed below.

The core disagreement between the reviewer and ourselves is on scope. Reviewer #2 requests additional experiments (prediction in ungauged basins, alternative parameterizations, multivariable constraints) that were not part of the original study we are analyzing. Our paper is explicitly a diagnostic analysis of results from "To bucket or not to bucket" (Acuña Espinoza et al., 2024), and while our work certainly opens up new avenues for future research, we do not see the benefit of extending the scope of this manuscript too widely - its value and beauty lie in the straightforward application of a single, new metric that allows for deep diagnostic insights with our proposed analysis routine. Handling too many deviations from this setting will reduce the clarity of the manuscript to an extent that hinders true understanding of our method instead of adding to it.

Based on the review comments, we have clarified our methodology to emphasize the idea that: identical LSTM architectures across all models enable an unbiased comparison using entropy, and we explicitly frame entropy as one diagnostic tool within a larger evaluation framework, not as a universal metric of physical adequacy. Moreover, their concerns about all parameters being dynamic have been addressed in a new section.

Reviewer 1 agrees most concerns were addressed but emphasizes the need to justify the exclusive focus on streamflow prediction, incorporate discussion of equifinality, and ensure entropy metrics are not biased by differing model architectures.

We thank the editor and Reviewer #1 for their constructive feedback. We have addressed all three remaining concerns in the revised manuscript as detailed in our responses below. We believe these revisions strengthen the manuscript's clarity and methodological rigor, and are confident that any doubts have been resolved.

Considering these responses, we respectfully submit that Reviewer #2's requested changes represent a substantial expansion of scope beyond what we believe is graspable in an already reasonably lengthy single manuscript. While we have addressed all points that are within the manuscript’s stated objectives and added a new section outside of them, implementing the remaining suggested changes would effectively require conducting a different study, moving away from our own objectives. We are further concerned that these changes might lead to an iterative cycle where new concerns arise, as the reviewer appears to envision a different study than what was originally in the preprint. Given that Reviewer #1 has recommended acceptance with minor revisions, that we have addressed Reviewer #2's concerns where feasible, and that some criticisms of Reviewer #2 are in contradiction to our results, we respectfully request that the editor evaluate whether the remaining points constitute revisions or represent a fundamental disagreement about research direction.

Kind regards,
Manuel Álvarez Chaves on behalf of the co-authors

Report #1 by Georgios Blougouras & Shijie Jiang (co-review team)

We thank the authors for responding to our comments. Most of our concerns have already been well addressed in the revised version of the manuscript. However, we believe that a handful of our aforementioned comments require additional revisions in order to ensure the quality of the manuscript.

We thank the co-review team for their feedback and address the remaining concerns in a revised version of the manuscript.

(following the labels of our initial review file)
General comment 2:
We still believe that there needs to be a clear explanation regarding why streamflow is the sole predictive focus of the manuscript. Streamflow (or any single-variable) prediction is not the only dominant application of HMs in the last years, and 'identifying correlations' (as the authors put it) does not paint the full picture; for example, current studies are actively using hybrid models to infer unseen variables and calibrate under a multi-task-learning approach (in an attempt to ensure process fidelity). We definitely agree with the added point in 1.4 (and the conclusions): by-passing the physical constraints indeed makes us concerned about the validity of inference. Yet, there is still no clear-in text clarification regarding why only single-output hybrid models are evaluated, as the scope of hybrid models can extend beyond single-variable performance. We believe this clarification would fit specifically in the Introduction, where the authors introduce the motivation for their study and their modeling choices.

Thank you for stressing this point. We recognize that hybrid modeling can extend beyond the prediction of single variables and the review process has made it clear that we should be more explicit about our rationale. We focus on predicting streamflow alone because we are evaluating the value of adding prior knowledge about the rainfall-runoff process in the form of conceptual models to an LSTM network. While we agree that if the goal is a hydrologically consistent hybrid model of high fidelity, multi-task-learning is the way to go, but it is by no means standard practice yet. Again, as highlighted in all versions of the manuscript and reiterated in the first revision, we are not promoting a specific approach to hybrid modeling, but offer an analysis for existing types of hybrid modeling that raise doubts - and this is exactly the setting of using conceptual models to apparently constrain the LSTM, for single-output prediction only. As suggested, we have clarified this motivation in the introduction, l. 154:
"Note that we focus on a typical single-task prediction (here: streamflow) to evaluate the value of adding prior process knowledge (here: rainfall-runoff) in the form of conceptual models to an LSTM network. Yet, we recognize the potential of hybrid models for multi-task learning, where models are evaluated on multiple objectives including multiple target variables, and anticipate that our proposed method can be readily extended to such evaluations in future work."

General comment 3:
Even though the authors aim to reserve the in-depth exploration of equifinality – entropy for a separate study, it nevertheless remains an issue that needs to be (at least partially) discussed,

as it is a core concern in hybrid modelling and hydrological models in general. The internal author responses to our comments are already a great step towards this direction - these (or equivalent explanations) should be added to the Introduction or the Discussion parts of the manuscript.

Thank you for this comment and it is reassuring to hear that our internal discussion is what you had in mind. We have now included the discussion of equifinality in the manuscript, both in section 3 relating to the synthetic examples (l. 495):

“Interestingly, equifinality did not pose an issue with synthetic data in our experiments, as all models achieved perfect predictive performance and the model was always identifiable under the right conditions. This matches the experience of Spieler et al. (2020). However, in real-world applications, equifinality is likely to be more pronounced due to measurement errors, incomplete observations of the system under study, and other sources of uncertainty. This issue is discussed further in Sect. 4.3.2.”

and in section 4 relating to the CAMELS-GB case study (l. 656):

“To return to the point of equifinality made in Sect. 3.4, as we have seen in this section, different hybrid model configurations may achieve similar predictive performance while exhibiting varying levels of entropy in the LSTM hidden state space and modifications to their internal behavior. We argue that high variability in parameter combinations represents an undesirable condition in terms of model structure specification. High entropy aligns with this perspective and, in general, entropy can be used to distinguish between equifinal models.”

Specific comment on section 2.3 ('I am concerned that the entropy metric... ML setup chosen in this study'):
Similar to the above point, we believe that the internal reply to the reviewer made by the authors should be incorporated in-text as well. The authors should instruct the readers on how to ensure that when they apply these entropy metrics themselves, these metrics do not end up being biased, e.g., from comparing models with different architectures.

Based on the reviewer's feedback, we have included a brief description of the setup in Sect. 3.2. More specifically, we provide more details regarding the choice of the ten hidden state LSTM and the number of basins (l. 324):

"The LSTM architecture of the baseline model and the hybrid models consists of ten hidden states. For our entropy analysis of the hidden states to be meaningful and fair, it is important to compare models of the same architecture. The choice of the number of hidden states was defined by the minimum required so that both the baseline model and hybrid models achieved equal performance. To aid in this process, the models were trained on a subset of five randomly selected basins (76005, 83004, 46008, 50008, and 96001) from the CAMELS-GB dataset.

(...)

We base our entropy analysis on equal performance, to ensure fair statements about the role of the conceptual component in a hybrid model ."

Report #2 by Anonymous Reviewer #2

I appreciate the authors' efforts in revising the manuscript and responding to the previous round of review. However, I think there are still some critical logical issues that remain unresolved:

First, in the synthetic case (Fig. 2), the central inference seems to be: conceptual model same as the "true model" or over-parameterized ⇒ low entropy; model different or underparameterized ⇒ high entropy. But this does not mean that models with high entropy are "wrong," nor that low entropy means "correct." Entropy is not a sufficient condition for the adequacy of a model structure. Over-parameterized models can have lower parameter entropies. Hydrologic models with more process representations may require a larger subset of parameters to be dynamic to accommodate their complex structure for good performance, whereas simpler models may need only a few dynamic parameters. Does this imply that the latter has a better physical representation?

In the synthetic case, indeed we found that overparametrized models had lower entropies but not lower than the "true" model. This is because for the overparametrized models there are degrees of freedom that the LSTM can adjust but which are not necessary.

We fundamentally disagree that "hydrologic models with more process representations may require a larger subset of parameters to be dynamic to accommodate their complex structure for good performance". Dynamic parameters, in our view, represent processes that one does not fully understand and therefore is letting a neural network fill this gap in understanding. If one could write a hydrological model using equations that describe each and every process in their full detail, there would be no need to couple this model with a neural network and therefore no entropy. The empirical results from the synthetic examples support this claim.

In addition, if only streamflow data from the "true model" is used to constrain the training (pretending this is the only information we know), we might think that Model 3 provides a better representation of hydrological processes since it has two-layer buckets (soil moisture and baseflow buckets) and gives an almost perfect streamflow prediction. However, when more information about the model structure or other variables is available, we see that this model is actually most different from the "true model." This simple example demonstrates why additional constraints are needed to diagnose whether a model reflects "bad" physics.

We respectfully disagree with this interpretation. To emphasize a point that we have argued before, we are evaluating the added benefit of a specific conceptual model in helping the hybrid model predict streamflow alone. For this specific example, without additional constraints, entropy shows that Model 3 does not provide a realistic interpretation of the synthetic truth (which has only one reservoir). As shown in Figure 4, Model 3 sits furthest to the right on the entropy axis relative to the data-driven model, whereas the true model is located furthest to the left. Further, the overparameterized models that can be reduced to the true model fall between the true model and the data-driven baseline. This positioning alone successfully diagnoses the inadequacy of Model 3's physical representation in characterizing the processes required to generate streamflow. This is precisely the diagnostic capability we propose. Our approach is fully consistent in interpretation and does not require "additional constraints" to "diagnose whether a model reflects 'bad' physics".

More rigorous experiments, e.g. adding additional constraints to the training (as suggested in the previous round of review), are required to demonstrate how entropy can be meaningfully connected with the physical representation of a conceptual model. The authors argue that “we're focusing on this relationship between model complexity and difficulty in prediction.” Can I then understand that entropy is related to model complexity, not to the correctness of the physical representation? However, what we need is not an easier model but a more physically correct model.

Both in the paper and in our previous answer we refer to the complexity of the LSTM measured by the entropy of its hidden states. If the conceptual model adequately represents the generation process of streamflow in the catchment, the LSTM will have low entropy/low complexity. If the conceptual model doesn’t accurately represent the streamflow generation process then the LSTM has high entropy/high complexity. Therefore it is the correctness of how the conceptual model represents the physical process.

Second, the entropy of LSTM states might not reflect parameter entropy. As noted in the previous review, we should not make all parameters dynamic. For the same model, the selection of dynamic parameters will change the entropy. In addition, the parameter ranges differ between models in the synthetic case, which might make comparisons and diagnosis problematic. A hybrid model with flexible parameter ranges can still simulate well even with the wrong structure, since the governing equations of buckets the simple conceptual rainfall–runoff models are similar to each other (summarized to Eq. 1).

We appreciate the reviewer's considerations regarding measuring entropy and model comparison.

As shown with the synthetic examples, the LSTM states are an unbiased way to analyze the entropy of the parameters and to compare between models with different numbers of parameters and different value ranges. Each hybrid model is coupled with an LSTM of identical architecture. While the conceptual models differ, they share this common LSTM component, which provides a consistent basis for comparing entropy across models.

Regarding dynamic parameter selection, this paper analyzes results from "To bucket or not to bucket" (Acuña Espinoza et al., 2024), which established the hybrid modeling framework with all parameters made dynamic. The current work focuses on diagnosing and interpreting those results through entropy analysis, rather than conducting new experiments with alternative architectures or parameterizations. Further, we do not see the fundamental difference to the approach we have proposed: if only specific parameters are made dynamic, then it is among these parameters to compensate for any misspecification in the conceptual constraint, so the individual entropies per parameter are expected to increase, and it is joint entropy over all hidden states that we measure, so we still see entropy as a fully consistent and fair metric.

To exemplify this, we have added a new Section 4.4 to the manuscript, with the previous Section 4.4 moved to Section 4.5. Subplot a) of the new Figure 12 shows the NSE CDF curves for the LSTM, SHM, Hybrid SHM, and a new model variant: SHM with only the parameter β in the soil moisture reservoir made dynamic (7 static parameters, 1 dynamic parameter). As expected, this single-parameter dynamic model shows a slight performance increase compared to the fully static conceptual model, but does not match the performance of the model with all dynamic parameters.

Subplot b) presents the entropy distributions for these models. The fully static conceptual model has no entropy and therefore is not shown in the figure. The single dynamic parameter variant exhibits an increase in entropy, even higher than the Hybrid SHM model with all parameters dynamic. This observation is consistent with our framework: when the LSTM must compensate for model misspecification through only one degree of freedom instead of eight, its activity (and thus entropy) increases substantially without proportional performance gains. To put it in different words, for SHM β dynamic, the LSTM has to "fix" the entire conceptual model through a single point of entry. By contrast, in the Hybrid SHM case where the LSTM is able to make smaller adjustments across multiple model components, this results in higher performance and lower entropy.

This example demonstrates that entropy serves as a diagnostic for how the neural network component compensates for structural inadequacies in the conceptual model. One can imagine additional experiments where parameters are selectively made dynamic or static to diagnose individual components in the conceptual model, with entropy guiding this process toward a representation that most accurately matches the natural system.

The reviewer correctly notes that parameter ranges differ between models. However, our entropy metric is derived from LSTM hidden states, not directly from parameter values. As explained previously, this provides a normalized comparison framework where the consistent LSTM architecture ensures comparability despite differences in the underlying conceptual models.

Third, the hybrid model using all static parameters could have similar performance to the one with some dynamic parameters in predictions for ungauged basins. I am not sure why the authors claim: “The potential overwriting of physics constraints happens during the training phase, and hence it is natural and logical to analyze it in the respective basins that these data are available for.” One can still calculate the entropy of the LSTM and parameters. In this case, can we say the model using all static parameters has a better physical representation than the one with dynamic parameters? However, they use the same physical model. How the NN part of the hybrid model is designed also matters. Again, equating low entropy with “adequate physics” and high entropy with “physics being ignored” oversimplifies the problem.

I think these issues, along with those raised in the previous round of review, need to be addressed in the current work; otherwise, the results could mislead readers. Unfortunately, the authors defer them to future work, which prevents me from supporting the publication of this manuscript in its current form.

We appreciate the reviewer's detailed concerns and address them systematically below.

Regarding predictions in ungauged basins (PUB), in our previous response, we focused on the fact that the overwriting is learned during training, so it makes sense to us to investigate this behavior for those basins with available data. However, we agree that one can *additionally* analyze the entropy when using the trained model to predict in an ungauged setting, to assess the amount of overwriting happening - it just doesn’t have the 1:1 translation of overwriting that *had* to happen to achieve good performance in fitting data, which is the core interest of our study.

Nevertheless, this paper is a follow-up to "To bucket or not to bucket" (Acuña Espinoza et al., 2024), which did not include PUB experiments, and we do not see the merit of extending our analysis to this setting because its interpretation does not add anything more didactic than what we have shown so far.

Outside of the ungauged setting, we do see the question of comparing models with dynamic versus static parameters as interesting. In fact, we observe a few instances where conceptual models with static parameters achieve performance equivalent to their dynamic counterparts. For example,

Basin 85001:

Model	NSE	Entropy
LSTM	0.921	-119.16
Hybrid SHM	0.923	-133.26
SHM	0.900	

Basin 94001:

Model	NSE	Entropy
LSTM	0.935	-121.26
Hybrid SHM	0.957	-128.56
SHM	0.926	

In these cases, the modest performance improvement of the hybrid model over the static version, combined with entropy values lower than the data-driven baseline, suggests adequate physical representation. The hybrid framework with dynamic parameters provides only marginal benefit for these particular basins. We allude to these specific examples in Section 4.3.1 and, more specifically, in Figure 7.

Concerning the influence of neural network architecture on entropy, we reiterate that all models in our comparison use identical LSTM architectures. This controlled setup allows us to attribute differences in entropy to the conceptual model component, as explained in our response to the previous point.

As a final point, we respectfully disagree with the reviewer's last statement saying that our interpretation oversimplifies the problem. Our framework interprets entropy as follows: low entropy indicates that the conceptual model provides an adequate physics-based representation that effectively constrains the hybrid model's behavior. High entropy indicates that the conceptual model misrepresents aspects of the natural system, requiring the neural network component to compensate through adjustments in the conceptual model's parameters to achieve accurate predictions. In such cases, the success of the hybrid model should be attributed primarily to the neural network's flexibility rather than to the physics-based conceptual structure. This diagnostic capability: distinguishing whether good performance stems from appropriate physical structure or from a neural network that compensates, is what entropy reveals and what we aim to communicate.

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

- Spieler, D., Mai, J., Craig, J. R., Tolson, B. A., & Schütze, N. (2020). Automatic Model Structure Identification for Conceptual Hydrologic Models. *Water Resources Research*, *56*(9), e2019WR027009. https://doi.org/10.1029/2019WR027009