Reply to RC2 'Comment on egusphere-2025-1699', Anonymous Referee #2

Review for HESS manuscript: When physics gets in the way: an entropy-based evaluation of conceptual constraints in hybrid hydrological models by Manuel Alvarez Chaves et al.

The topic and method discussed and used in this manuscript are interesting and timely—when and how the addition of physical principles genuinely enhances model performance and improves the representation of underlying physical processes. Overall, I am sensing some logical issues, making the findings might not be conclusive or fully compelling. They designed a synthetic case and a real application on CAMELS-GB dataset. In the synthetic case, they found that the LSTM is the dominant factor in the hybrid model and can even effectively adjust a nonsense formulation to perform as well as the other models. The LSTM (for parameter estimation in the hybrid model) applied for more accurate process-based model structure has less entropy. The nonsense model has the highest entropy. In the CAMELS-GB application, LSTM has the lowest entropy and highest performance, followed by the nonsense model structure with the second lowest entropy and the hybrid bucket with the second highest performance. With that, they made the conclusion that the connected physical model did not simplify the prediction nor improve performance.

We sincerely would like to thank the reviewer for their evaluation of our work and the time they have taken to engage with our manuscript, and we appreciate the interest in advancing hybrid modeling.

First we would like to clarify some points particularly for the synthetic case study as described by the reviewer. In the synthetic case study, we did not propose an architecture that was "nonsense" but rather conducted a controlled experiment where we generated synthetic data using a known "true" conceptual model, then tested eight different hybrid models as working hypotheses to see how well they captured the underlying data-generating process. When evaluated solely on their ability to fit the data, all eight hypotheses performed equally well in terms of MSE. However, we then applied our proposed entropy measurement to assess whether performance was achieved through the conceptual model component or through compensation by the LSTM component modifying the conceptual model's parameters.

Our key finding was that the hybrid model containing the "true" conceptual model required the least assistance from the LSTM (lowest entropy), while models with architectures very different from the true underlying process required the most assistance from the LSTM (highest entropy). This demonstrates that our entropy metric can distinguish between conceptual models that perform well for the right reasons versus those that achieve good performance just through compensation by a data-driven component, while traditional performance metrics cannot make this distinction.

Regarding the CAMELS-GB case study, you provide an accurate description of our entropy rankings, yet we want to emphasize a very important distinction: entropy is not equivalent to performance. While entropy measures how much the LSTM component must compensate for the conceptual model, model selection should still consider traditional performance metrics like NSE. In our comparison across specific basins, we intentionally selected examples where all models achieved similar performance to match our synthetic case findings and focus on the insights which entropy can reveal.

It is not our intention to make definitive statements about overall model choice, and we will carefully go through a revised version of the manuscript to avoid this impression. The primary focus of this paper is on a novel model evaluation and comparison method that adds insights to better understand why specific hybrid versions perform well or not so well, rather than guiding a process of model selection. We will highlight this in our revision to make sure that our scope and conclusions are presented more clearly. We also go into more detailed specifics by addressing each of the points made in this review.

Here are my major thoughts on hybrid models, the experiments designed in this work, and the conclusions drawn from it

1. The LSTM in the hybrid model is used for parameter estimation, while the simulation is conducted by the process-based model (PBM). The main reason for incorporating PBMs is to improve model interpretability and performance, especially when data is limited (by reducing the searching space). The latter is achieved through physical constraints provided by PBMs. The physical constraints include not only mass balance for each bucket but also the interactions between fluxes. More specifically, the hybrid model provides clearer expressions (better interpretability) of the physical processes than black-box models (e.g., LSTM) and enables the calculation of internal variables. This means that we can incorporate more data as target values to further constrain or evaluate the model.

Many studies, including those by the authors, showed that the internal variables such as soil moisture and ET are comparable to observations. If the physics are totally incorrect (or, as the authors commented --- “physics-ignored”), how is it possible that these internal variables make sense? If some of the modules carry realism to some extent (which would make “physics-ignored” incorrect), can the authors’ method ascertain what is good or correct? If the method cannot ascertain what is good, how can it separate the good from the bad?

Thank you for this comment. We would like to clarify that we use "ignored" rather than "incorrect" physics because this better describes what may happen in hybrid models. Physical laws such as conservation of mass and physical principles like the interaction of fluxes are indeed encoded in the model, but in some cases they might not be satisfied in the ways in which the modeller might expect.

If the prescribed constraints are not respected in the hybrid model, then also interpretability is not given as such. This emphasizes the need for and usefulness of our proposed method: we need to better understand if the hybrid model under investigation truly respects the physics as intended (in that case, interpretation and consistency tests make sense and are meaningful), or if the data-driven component effectively changes the structure - up to essentially manipulating all processes, leaving only mass balance as effective physics constraint (in that case, interpretation based on the initially prescribed physical processes would be misleading and not suitable for advancing science).

We suspect that hybrid models, more often than not, show an effective structure that is notably different from the initially prescribed physics constraint. We hope to motivate the readership and community to perform extensive tests with our proposed methodology to confirm or dismiss this hypothesis.

The reason why so far this issue has not been widely documented might lie in the fact that even hybrid models that are constrained by “inadequate physics” have the capacity to show high correlations with unobserved variables. Our Hybrid Nonsense model demonstrates this clearly as was shown in Acuña Espinoza et al. (2024). Despite deliberately encoding nonsensical interactions in between its fluxes and storages, this model still achieved a correlation with soil moisture (from ERA5-Land data) of 0.85 that is nearly identical to the more physically reasonable Hybrid SHM with a correlation of 0.86. This shows that compensation by a neural network can make even poorly encoded knowledge of physics appear to work well. One possible explanation is that soil moisture dynamics are mostly driven by boundary conditions (rainfall - soil moisture increases, no rainfall - soil moisture decreases), making it not a particularly challenging test. It remains an open question how consistency checks should be designed to be truly informative in a physics context; our proposed method will help establish the ground for interpretation of such consistency checks by warning the modeler if the prescribed constraints are actively compensated for.

The key contribution of our work is hence providing a systematic way to distinguish when physics are consistently utilized versus ignored and manipulated. Our entropy metric, benchmarked against a purely data-driven baseline, enables this separation. Higher entropy than the baseline indicates "bad” physics, i.e. constraints that need compensation, while lower entropy indicates "good” physics, i.e. contributing meaningfully to model performance. “Good” and “bad” is used as abbreviation for these longer definitions in the manuscript and properly explained as such; if that is perceived as too bold, however, we would consider replacing those terms in the caption of Fig. 4.

Currently, the model in this study is trained only with streamflow data and evaluated against streamflow. However, the authors could design a more complex synthetic case with additional internal variables such as ET, snow, soil moisture, and baseflow, and explore the following:

a. If we add more constraints to the internal variables (by using them as additional targets), can the nonsense formulation still achieve the same performance as the correct model?
b. Can a model with a nonsense formulation provide accurate simulations of these internal variables when trained only on streamflow?
   In real cases, even when we lack direct observations for some variables, we can still compare model outputs to satellite or reanalysis data to gain insights into the accuracy of the internal variables and model structure.

c. When your hybrid models produce accurate predictions for both internal variables and streamflow, do you still observe an increase in entropy when adding physical models? Or, if entropy increases, can you still maintain the same level of accuracy?

Thank you for these suggestions. These are indeed questions we have considered ourselves and would be valuable to explore in future work. However, we believe it's important to maintain a focused scope for this study to ensure we provide a thorough answer to our central research question. The performance of hybrid models as such, and the best training strategy (single target vs. multi-objective) to increase it, are not of core interest here (as we noted above, the entropy metric is not to be confused with a performance score).

Furthermore, we do not wish this study and publication to become a comprehensive evaluation of the Nonsense model. The motivation for this study arose from the observation in Acuña Espinoza et al. (2024) where good predictions of streamflow were achieved without the need to add a proper physics-based model like SHM, but using Nonsense instead. This led us to investigate why you could seemingly add any conceptual model structure, even deliberately "nonsensical" ones, and still achieve equally good hybrid predictions of streamflow. We believe our current scope allows us to provide a solid understanding of this phenomenon through both our synthetic case study and the CAMELS-GB case study, and we see the questions you've raised as natural and important extensions for future research.

Regarding the broader question of using correlations with internal variables to validate hybrid models: as discussed above, there is a need to increase the rigor of such analyses by putting the correlation results into the proper context. While important hybrid modeling studies have used correlation of internal variables as validation (Feng et al., 2022), from the other perspective, it has been shown that LSTM cell states can also achieve strong correlations with observed variables, even with individual cells presumably learning to track specific processes like snow accumulation (Kratzert et al., 2019).

Instead, we first need to understand what the hybrid model is doing, before we can interpret correlations. It might be that the internal functioning of the hybrid model is substantially different from what the modeler had put in, and hence a surprisingly high correlation with variables not used for training might just mean that the hybrid model found its way into an effective time-lagged representation of how rainfall increases the level of a storage, before producing discharge: this is what we observed in our test case with the Nonsense model. We therefore see our proposed analysis as a first and necessary step to put any correlation analysis on solid grounds. The result of a correlation analysis, however, will not have an impact on our method in any sense, which is why we don't see merit in extending the scope of the current manuscript to internal variables. Also, the question how we can better enforce physics constraints to prevent them from being overwritten is beyond the scope of this study; rather, the proposed method provides a tool for modelers to test the effectiveness of their way of constraining hybrid models.

2. Should we make all parameters dynamic? Similar to traditional conceptual models, hybrid models also suffer from overfitting when too many parameters are calibrated (or "learned," in the

case of hybrid models), especially when done dynamically. We should not make all parameters dynamic. This caution has been raised in several recent works on hybrid modeling. By making all parameters dynamic, we give the LSTM too much freedom and intentionally allow it to dominate the hybrid model. One key point I want to emphasize is that LSTMs (and neural networks in general) are "lazy" models—they tend to find the most convenient (easiest) way to make predictions. Rather than learning the true functional relationships of parameters, they often memorize patterns in the time series, and often special effort is needed to make it extract the necessary information. To discourage this "laziness" in neural networks, two approaches can be considered:

a. In addition to the temporal test as done in this work, prediction in ungauged basins or regions is a useful evaluation method. Verifying the accuracy of internal variables is another effective strategy. If a hybrid model demonstrates good spatial generalizability, how does its entropy compare to that of a pure LSTM?
b. If we only allow a limited number of parameters to be dynamic, for example, just beta (shape coefficient in the soil moisture zone), can all models still achieve the same level of performance? How does their entropy compare to that of the pure LSTM?

The authors touch on a similar point in Lines 645–648: “Notably, this process occurs entirely under the hood. If we had evaluated performance using only NSE, we might have mistakenly concluded that the Nonsense constraint was just as valid as SHM or Bucket, since all three achieved the same performance when paired with the LSTM.” I hope the authors can consider my suggestions in their entropy analysis experiments. The current experiments and narrative might unintentionally lead readers to conclude that the hybrid model lacks of effectiveness and interpretability.

Thank you for these suggestions. In fact, ultimately we wish to help improve the effectiveness and interpretability of hybrid models, by providing a rigorous analysis tool that helps looking under the hood. We challenge the assumption that hybrid models become effective simply because we add a specific structure like SHM. As reflected in the lines you highlighted, our intention is to demonstrate that, when predicting streamflow alone, some models might be good for the wrong reasons, and we aim to provide tools to identify such cases, to have a chance to correct them and develop better ways of building hybrid models.

We agree that not all parameters need to be dynamic, and our implementation actually provides the flexibility to keep certain parameters static while making others dynamic. Your suggestions represent excellent directions for future research. For example, if we freeze certain parameters, does the LSTM increase its reliance on remaining parameters (increasing entropy), or does it find more elegant solutions with the remaining degrees of freedom (reducing entropy)? These questions exemplify how this study opens new research avenues while advancing model evaluation beyond simple predictive performance to include model complexity, a topic that is typically not addressed.

Nevertheless, for this study we deliberately followed that of Acuña Espinoza et al. (2024), as we specifically wanted to address the questions those authors raised. To emphasize, we didn’t do any new modeling for this study but analyzed previously published results.

With respect to prediction in ungauged basins: we focused on temporal rather than spatial extrapolation because we also followed the scope of Lees et al. (2021), which matches the standard benchmarking practice for CAMELS datasets. Theory-wise, we do not see the direct benefit of applying our entropy metric to the predictions of hybrid models in ungauged basins. 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. In a secondary step, one could of course also use our method to analyze the variability of LSTM hidden states when predicting in ungauged basins, but the interpretation becomes more difficult - we see this as an open question for future work.

Regarding your point about neural networks being "lazy" models, please indulge us as we'd like to offer a philosophical perspective. We believe this apparent "laziness" sometimes aligns beautifully with the principle of Occam's razor in that what appears to be lazy may actually represent profound insight that recognizes that a complex problem can have a surprisingly simple solution. In our case, the LSTM's ability to ignore unnecessary complexity in models like SHM and focus on what's truly essential for streamflow prediction could be seen as elegant, efficient and parsimonious problem-solving rather than laziness. The fact that the LSTM could transform even our Nonsense model into an effective predictor suggests it identified the core requirements for this prediction task, potentially revealing important insights about what's actually necessary for predicting streamflow accurately in this dataset. Further, we designed the didactic synthetic examples deliberately such that they could reveal if LSTMs tend to do “overcomplicated” things or fail by only memorizing time patterns. It was reassuring to see that they identified the simplest and truest representation possible.

3. I also felt the entropy argument has some logical issues: The fact that LSTM has the lowest entropy simply means it has no reason to care about processes. Let’s think of process-based Earth System models with land surface processes like energy balance and carbon cycles, with many complicated calculations and quite likely lower NSE values. Now, let’s say we replace a component with large variability but minor influence on streamflow with an NN-based dynamic parameter. The dynamic parameter will learn some of the missing dynamics but will generate a large entropy with a quite small gain on streamflow performance, but does this mean it is worse than a model that completely ignores these dynamics? Moving one step further, if one “downstream component” to the NN has large structural error and generates lots of noise, the NN ends up ignoring this component (giving it nearly 0 weight), and route information through other paths. The total entropy may end up being lower. Does this mean the second one is a more realistic model?

Thank you for raising these questions.

To reiterate, our measure of entropy is designed to evaluate whether added model components contribute meaningfully to make the task of prediction easier or whether the neural network must compensate for their inadequacy. Entropy is used as a diagnostic tool. When we argue that entropy relates to the challenge of prediction (lines 394-401), we're focusing on this relationship between model complexity and difficulty in prediction. Further, as stated above, we do not wish to provide a model selection method that pits performance against entropy. Rather, we provide a tool that transparently shows how much the LSTM has to compensate, which is expected to be very valuable in model evaluation. We do not see the point of providing a universal recipe of how to balance this with performance. The main driver of advances is, from our perspective, understanding why our models perform as they do.

So coming back to your specific example, we believe that if a hybrid model learns to bypass structurally flawed components, this provides valuable diagnostic information. While this doesn't make the model more "realistic" in terms of process representation, it reveals which processes actually contribute to predictive skill versus which introduce unhelpful complexity. Understanding when and why processes are ignored can guide us toward better representations or help recognize when we're asking models to represent processes they cannot adequately constrain with available data.

4. One should not think that the NNs in a hybrid model learn only true physics --- it is a mixture of true missing (to be learned) processes and compensation for structural, parametric and even numerical deficiencies. The baked-in assumptions serve as constraints to limit the searchable subspace, hopefully making the framework more generalizable and giving meaning to intermediate variables. The hope is that the signal overwhelms the noise or that we can extract useful insights from learning. We verify the meaning using additional observations.

My point is that one should base such analysis on a case-by-case basis and the conclusion obtained by the experimental design in this study may not generalize to another case. Entropy measures variability, not correctness. A model may need high variability in parameters due to data noise, poor observability, or inherent system variability—not necessarily because the constraints are wrong. The authors acknowledge uncertainty later (Sect. 4.3.2 and 4.5) but still lean heavily on this deterministic interpretation throughout. Equating low entropy with model adequacy and high entropy with “physics being ignored” oversimplifies complex interactions between model structure, data, and learning dynamics.

Thank you for this comment. In rainfall-runoff modeling, where we work with significant abstractions of real-world dynamics, we acknowledge that what we learn are not "true physics" but rather macro-scale insights about system behavior, no matter if we talk about conceptual or hybrid models. Further, here the NN is not used to learn “true” physics, as these are assumed to be encoded by the conceptual model which serves as the head layer of the LSTM. Rather, the LSTM learns how to bias-correct the conceptual constraint to arrive at close-to-true physics.

As the Reviewer states, models will struggle with data noise, poor observability, or unresolved system variability, and this will have an impact on how much effort the LSTM has to put in to make the rigid conceptual constraint perform well under these conditions. Hence, while it will be difficult to interpret individual entropy values as “high” or “low”, the comparison with the reference entropy value of the pure LSTM, and potentially other hybrid model structures as done in our study, for a given catchment is highly informative. This is why we promote constructing at least parts of the model evaluation axis in Fig. 4, and why we look at basin-specific rankings in Fig. 11 to complement the results from simply comparing entropy values across basins in Fig. 10. It will be interesting indeed to further refine our understanding how “adversities” in the data of the basin to be modeled impact the difficulty of the modeling task and hence the entropy metric in future work.

We agree that evaluation should be case-by-case, which is why we promote a widely applicable method, not its analysis outcomes. We made a careful effort to both provide an overview of results for that specific large-sample data set, and additionally investigated specific basins in-depth for our analysis in Sections 4.2, 4.3.1, and 4.3.2. The purpose of these analyses is to let the readers gain intuition about the possible outcomes of our analysis method and their interpretation, such that they can apply the method to their cases of interest. It will be valuable to find out if our results generalize across many different large-sample datasets, and how they depend on the specific hybrid model architecture. Again, with this study, we hope to stimulate further research along these lines, to advance hybrid modeling in hydrology but as Reviewer 1 pointed out, as well in many other disciplines.

5. Ignoring causality. None of the analyzed information metrics measure causality, while causality is a key value brought in with the incorporation of physics. The physical component can ensure that higher temperature can lead to higher evaporation, or ensuring that precipitation (not humidity) drives runoff. The causality and baked-in sensitivity allow future climate projections to be more reliable. Your information metrics do not reflect such guarantees.

Thank you for this observation. You are correct that our entropy-based analysis does not measure causality, because it is not designed to do so, and we also do not claim anywhere in the text that it would. We do agree that what the Reviewer describes as causality is a desirable property and insightful methods should be developed to measure/detect it in hybrid models. This is an open question, yet unrelated to the topic and goal of our study.

Minor comments:

Line 102–104: There are several other studies that have compared hybrid models with LSTM. These could also be mentioned here for completeness.

Thank you for this suggestion. Indeed there are, and we will include some additional references to comparisons between hybrid models and LSTMs in Section 1.3 of a revised version of the manuscript, while keeping these lines focused on motivating the current study.

Line 451–453: This depends on which aspect of the problem you're examining. Hybrid models should be compared not only to LSTM but also to traditional process-based models. Given the same structure of a process-based model, the hybrid version can yield better performance. Compared to LSTM, hybrid models offer better interpretability. The fact that a hybrid model can make a nonsensical model perform well does not mean that a hybrid model built with a correct process-based structure is invalid. As I mentioned in my major comments, the model should be evaluated more comprehensively. Previous experience and domain expertise can help reduce equifinality problems and prevent obviously nonsensical model configurations.

Thank you for this comment. We agree that hybrid models may provide valuable capabilities between traditional and data-driven approaches, and we are not saying that well-designed hybrid models are invalid. As we have pointed out before, it is critical to understand if the prescribed physics are obeyed or not, before making a claim about interpretability.. The fact that our "nonsensical" structure could be transformed into an effective predictor suggests we should pay more attention to what is effectively happening in the hybrid model structure, and be more critical about whether what we're adding actually contributes to the task at hand. Our entropy analysis is intended as a complementary tool to help identify when neural networks are compensating for structural choices, allowing more informed decisions about the merits (such as interpretability) of a model.

Line 561–562: This supports my view that LSTM is a “lazy” model. It tends to find ways to bypass physical constraints, especially when dynamic parameters are allowed.

While we wouldn’t necessarily agree with the notion of “lazy” (we rather find it smart and reassuring that the LSTM finds a parsimonious, efficient and skillful representation of the system), we fully agree that the LSTM finds ways to bypass physical constraints if they are not helpful for the prediction task, and this is exactly what we shed light on with our method - in a quantitative way by means of our entropy metric, and qualitatively by analyzing the overwriting behavior of the dynamic parameters.

Line 637–638: A simplified task means that LSTM has found an easier path to make predictions, but this does not necessarily indicate higher accuracy. Multiple evaluation metrics should be used to support conclusions

We disagree with this point. All models were trained and evaluated using the same loss function (NSE or MSE), ensuring accuracy is directly comparable. So the LSTM indeed achieved higher accuracy in streamflow prediction. What we refer to as a "simplified task" is prediction with lower entropy: if the physics constraint took over main parts of prediction, the hidden states in the LSTM wouldn’t have to vary as much as in the pure LSTM, and hence entropy would be lower. While additional metrics could provide more detail and further insight, our study focuses specifically on the “effort” of the LSTM in this hybrid model architecture.

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
- Feng, D., Liu, J., Lawson, K., & Shen, C. (2022). Differentiable, Learnable, Regionalized Process‐Based Models With Multiphysical Outputs can Approach State‐Of‐The‐Art Hydrologic Prediction Accuracy. *Water Resources Research, 58*(10), e2022WR032404. https://doi.org/10.1029/2022WR032404
- Kratzert, F., Herrnegger, M., Klotz, D., Hochreiter, S., & Klambauer, G. (2019). NeuralHydrology—Interpreting LSTMs in Hydrology. *arXiv:1903.07903* [Physics, Stat], 11700, 347–362. https://doi.org/10.1007/978-3-030-28954-6_19
- Lees, T., Buechel, M., Anderson, B., Slater, L., Reece, S., Coxon, G., & Dadson, S. J. (2021). Benchmarking data-driven rainfall–runoff models in Great Britain: A comparison of long short-term memory (LSTM)-based models with four lumped conceptual models. *Hydrology and Earth System Sciences, 25*(10), 5517–5534. https://doi.org/10.5194/hess-25-5517-2021