

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.

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?

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?

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.

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?

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.

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.

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.

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.

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.

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