Summary of the Revised Version

We sincerely thank the referees for their careful and constructive comments. We have substantially revised the manuscript to clarify the scope of the temporal-difference loss regularization (TDLR) framework, improve its logical flow, make the benchmark comparison more transparent, and more carefully distinguish model-derived functional representations from physically verified hydrological processes.

The major revisions are summarized as follows:

1. Reframing of process identification and identifiability

   We have revised the manuscript throughout to clarify that TDLR does not establish physically unique process functions from streamflow data alone. The extracted representations are now explicitly presented as robust and testable mechanistic hypotheses, conditional on the prescribed bucket structure, mass-balance constraints, input data, and output-range control. We further distinguish optimization/re-extraction stability from physical identifiability. The conclusion now explicitly states that the true catchment-scale response functions cannot be uniquely determined from the available observations.

2. Clarification of the role of TDLR in the closure problem.

   We reframed TDLR from solving or uniquely identifying the closure problem to conditionally searching for admissible catchment-scale functional representations as a hypothesis-generation step. The revised framework emphasizes an iterative workflow in which candidate representations are evaluated against hydrograph reconstruction, mass balance, repeated extraction, independent observations, and catchment information.

3. Improved benchmark fairness and information-use comparison.

   We added pure LSTM benchmarks both with and without observed Qt−1, allowing a direct comparison between TDLR and LSTM under the same information setting. We also added a fully recursive/free-running TDLR experiment using estimated Qt−1. The revised Table 1 reports both NSE and KGE, and the related discussion is consolidated in Sect. 4.1.

4. More cautious interpretation of the mass-balance residual.

   We now explicitly describe Lt as a residual that may contain interbasin groundwater exchange, measurement error, and structural mismatch. We added a ±20% precipitation sensitivity analysis and explicitly state that the observed residual cannot uniquely distinguish inter-basin groundwater flow (IGF) from persistent measurement bias or structural error. We therefore present IGF as a mechanistic hypothesis rather than an identified process and specify the independent observations required for direct testing.

5. Correct the Jiji catchment delineation and retrain all Jiji models.

While re-examining the Jiji dam catchment maps and information, we found that the original delineation and the corresponding meteorological forcing dataset omitted one headwater subcatchment. We corrected the delineation, re-collected the input data, and retrained all deep learning-related models and frameworks for Jiji. We substantially revised the Jiji results across all figures and tables in the revised manuscript. Correcting the delineation of Jiji and retraining all models improved hydrograph reconstruction accuracy across all Jiji models, but did not reverse the performance ranking in Table 1 or the main conclusions. The qualitative patterns of the extracted representations at Jiji are consistent with the original version.

6. Redraw, add, and rearrange the figures, tables, and sections.

a. Figure 1 has been redrawn to add the geological condition.

b. Figure 4 has been redrawn to add KGE and align the presented period.

c. Adding the free-run TDLR with estimated Qt−1 and KGE metric in Table 1 and rearranging Table 1 and related discussion in Sect. 4.1.

d. Fig 5(a) and (b) in Sect.4.2.1 are from the original Fig 5(f), and the new comparison between the external AET product and calculated AET is Fig 5(b).

e. To discuss canopy storage dynamics in Sect. 4.2.2, Fig 6(a), (b), and (c) are from the original Fig 5(e), Fig 1(d), and Fig B1

f. In Sect. 4.2.3, Fig 7(a), (b), and © are the original Fig 5 (c), Fig 5(d), and Fig 1 (c) to discuss the surface water routing.

g. In Sect. 4.2.4, Fig 8(a), (b), and (c) are the original Fig 5(a), Fig 5(b), and Fig 1(b) for runoff-generation and subsurface storage dynamics discussion.

h. In Sect. 4.2.5, Fig. 9 is the original Fig. 7. It shows the functional representation comparison between the ones extracted from TDLR and the dPL-HBV model.

i. In the revised manuscript, we added the 95% confidence interval of the curve fitting in Fig 5(a), Fig. 6(a), Fig 7(a) and (b), Fig. 8(a) and (b).

j. In Sect. 4.3, we added the sensitivity test for Lt in Fig. 10 to show that the residual signal is still significant even considering the precipitation measurement error, supporting the hypothesis generation for IGF.

k. Added Fig 12 in Sect. 4.5; it presents the workflow between AI and export collaboration with mechanistic hypotheses generation with TDLR. In the added Sect. 4.5, we discussed the concept of the extracted functional representation treated as data-driven mechanistic hypotheses generation.

Response to Referee 1

General Comment

Liu and colleagues provide an interesting piece of work, where they explore the possibility of extracting interpretable representation of catchment hydrological processes using deep learning methods. The topic has been highly explored recently and it is worthy the study,

but I have some major issues that need to be addressed before the paper might be considered for publication at HESS.

Major comments:

1. *The paper is somewhat dense in the methods part. The authors could try to synthesize the most relevant information so that the reading could flow more naturally.*

Response: We thank the referee for the constructive feedback on the manuscript's readability. We have restructured and reorganized Sections 2 and 4 to improve the flow of the presentation. Specifically, we (1) reorganized the description of the study area, (including its figures) and Sections 4.1-4.2, Figure 5; (2) reordered the material so that each interpretable function is introduced together with its independent supporting evidence and information and placed near the corresponding results; (3) removed inferential and discussion-type content from the Method section, relocating it to the Results and Discussion; and (4) added Section 4.5 present the concept of data-driven hypothesis generation. Regarding the flow of the revised Section 4.2, we started with the figure and the result first, then presented the independent information related to the result, and lastly provided the interpretation, theory, and mechanistic hypothesis. All changes are marked in red in the attached file. Additionally, we deducted the Sect. 3.1 in revised manuscript to make the idea of methodology more concise.

2. *In many parts, the authors start with some discussion or even some results, without stating the relevance before. The readers should be aware of why this is relevant before reading, or the text should be appealing anyway. In the current form, it makes the reading harder. I suggest that the authors review the manuscript to make sure that there is a good balance and flow between relevance and what is written. This could significantly improve the manuscript. Some examples are:*

    1. *In L145-148, the relevance of all the discussion about precipitation is put just after.*
    2. *Paragraph around L446: Again, please state the relevance before so the readers knows why does it matter.*

Response: Thanks to the referee for pointing this out. In the original manuscript, the discussion content, especially in Sections 2.2 and 4.2, was separated from the relevant results and motivation. To address this, we merged Sections 2.2 and 4.2, so that the motivation and relevance are now presented before the corresponding discussion, as also described in our response to comment 1 above. In addition, we revised the specific passages the referee identified: at Lines 145–148, the relevance of the precipitation discussion is now stated at the beginning of the paragraph; and the paragraph around Line 446 now opens with a statement of why the analysis matters before the results are presented.

3. The authors use in their model Qt-1 as input for the model. I have no issues with that, but during the comparison with the hybrid model, the comparison becomes unfair. The authors should either compare to a version using such an input as it was done for the LSTM, or acknowledge that their comparison about realism is somehow more complex than what is shown (Section 4.4.).

Response: We thank the referee for pointing out this issue. We rewrote Section 4.4 and combined the context in Section 4.1. To make the comparison fairer and highlight our main argument that fixed prior functional forms might limit deep learning's ability, we reframed the comparison into two groups, defined by whether previously observed streamflow is used. The first group is hydrograph reconstruction with previously observed streamflow, TDLR with observed Qt-1 , and Pure LSTM with Qt-1 added as input. The second one is dPL-HBV (static/dynamic) and Pure LSTM (64/365 days). We added the KGE and free-run TDLR with estimated Qt-1, as shown in Table 1. Comparing the reconstruction accuracy between the two types indicates whether adding prior knowledge and concepts in our study area causes an information-usage bottleneck. For the more detailed discussion, please see the revised manuscript from lines 295 to 329.

Table 1. Comparison of simulation accuracy among the proposed TDLR framework, dPL-HBV, and the pure LSTM. We used Nash-Sutcliffe efficiency (NSE) (Nash and Sutcliffe, 1970) and Kling-Gupta efficiency (KGE) (Gupta et al., 2009; Knoben et al., 2019) as metrics to assess simulation accuracy over the entire record. Bold values show the TDLR/model computes current streamflow with the previous observed streamflow, Qt-1, as input; All other models simulate the current streamflow from meteorological driving forces only. Numbers in parentheses (64/365) denote the LSTM input sequence length in days. Note that dPL-HBV, as a storage-based model, cannot ingest observed Q(t−1); comparisons are therefore most meaningful within each group.

	TDLR with observed Qt-1	*Pure LSTM added Qt-1 as input (64/365 days)*	*Free run TDLR with estimated Qt-1*	*dPL-HBV (static/dynamic)*	*Pure LSTM (64/365 days)*
Shihmen (NSE)	**0.863**	**0.86/0.85**	0.807	0.833 / 0.802	0.84/0.86
Jiji (NSE)	**0.867**	**0.86/0.87**	0.612	0.611 / 0.676	0.81/0.81
Deji (NSE)	**0.823**	**0.82/0.84**	0.711	0.709 / 0.656	0.74/0.72
Shihmen (KGE)	**0.930**	**0.82/0.86**	0.867	0.860 / 0.857	0.89/0.91
Jiji (KGE)	**0.931**	**0.88/0.90**	0.801	0.688 / 0.800	0.80/0.90
Deji	**0.898**	**0.79/0.84**	0.826	0.701 / 0.728	0.78/0.76

(KGE)

4. Were the models regional LSTM (or regional hybrid) for the benchmarks? Also, was the model employed in this study a regional, or a model per basin? I ask because there is considerable literature about the use of single models for prediction tasks when dealing with LSTMs, and acknowledging this is minimally needed when making comparisons about model performances. For the sake of the study, any comparison is worthy, but if one claims practical consequences, one must be more careful.

Response:

We thank the referee for highlighting this important distinction. Our experiments are catchment-specific rather than regional. A separate model was trained for each of the three catchments, and observations or model parameters were not pooled across catchments. This applies consistently to TDLR, pure LSTM, and dPL-HBV. We have clarified this explicitly in Sect. 2 and Sect. 3.3. We also now frame the study as a catchment-scale proof of concept, rather than making claims about the practical superiority of TDLR for regional prediction. The cited large-sample/regional LSTM literature is acknowledged as an important complementary modeling paradigm; our objective is different: to examine whether functional representations can be extracted conditionally within an individual catchment. Therefore, per-catchment training is a natural setting. Hence, the concept of proof that uses deep learning to support hypothesis generation has been demonstrated in the revised manuscript. Extending TDLR to train the regional LSTM on a larger dataset and examining whether regionally trained models yield transferable or catchment-clustered functional representations is identified as future work.

Meanwhile, in pratical usage, the TDLR with previously estimated streamflow also added in Table. 1 in revised manuscript and previous response.

Minor comments:

L58-60: Did these authors say that?

Response: Thanks to the referee for pointing out the mistake. The cited authors do not discuss advanced parameterization but rather fixed functional forms for catchment-scale hydrological processes. We have accordingly rewritten the sentence, removing the reference to advanced deep learning parameterization. The revised sentence in Lines 57 and 58 is “However, there is a lack of clarity regarding the identification of catchment-scale prior functional forms for each hydrological process (Beven, 2006b; Blöschl and Sivapalan, 2006; Kirchner, 2006; Peters-Lidard et al., 2017).”.

L75: perhaps the authors meant: “there is currently” instead of “there was”?

Response: Yes. We thank the referee for this correction. We have revised the sentence from lines 77 to 79: " Nevertheless, there is currently no established method for deriving interpretable, stable functional representations from observational data as mechanistic hypotheses describing catchment-scale hydrological processes. "

Figure 4: Why use Qtrue and in the legend inside the figure use Qobs? Please stick to one.

Response: We thank the referee for pointing out the mistake in Figure 4. We have fixed the issue.

L341: Competitive to what?A benchmark? Some NSE literature threshold?

Response: We agree that the term competitive is not appropriately used. We have removed this term and rewritten the relevant text. Please note that the results and discussion about the benchmark comparison have been substantially revised.

There are many places where some claims, which are backed by literature or by the results themselves, are made, but such references are never exposed. Please make sure that you do not have any hanging claim. reference the paper that backed you up, or the figure. Some examples:

1. *L345: Where can we see this? Methods and Data? Please add the reference.*
2. *L385: where is such a covariance shown?*
3. *Paragraph L400: Your study area description talks about lateral flows and faults and gives some references. Why don't you do this here?*

Response: We agree that several claims in the original manuscript were not explicitly connected to their supporting evidence. We have reorganized Sections 2 and 4.2 so that the chain from results to independent information to hypothesis and theory flows more naturally, and we have checked the full manuscript for unsupported claims. Regarding the specific examples: in r

(1) Thanks to the referee for pointing out the relationship between memory effect and LSTM is not presented and providing the reference in the manuscript. To prevent the misunderstanding, we removed the sentence.

(2) The covariations between wetting area expansion along with the catchment precipitation grow is at Fig. 1(b) in the original manuscript. To improve the flow, we reorganized the order of the figures.

(3) We now support this paragraph with the same geological references used in the study area description with the revised Fig. 1.

L361: Word competitive used again.

Response: Corrected in the same manner as our response to the comment above; the term has been removed throughout the manuscript.

Figure 5: Please be verbose about what the y-axis variable means. This can help the readers to map directly to the interpretations without the need to come back to where they really mean.

L409: "(ii) smooth…" and "(iii) increasing first, then…" I have difficulties seeing it in the figure. Can you confirm that this is what you meant? Also reference which catchment each of these patterns refer to exactly.

L509: which plateau? I have difficulties to see a plateau from the figure.

Response: We thank the referee for pointing out these presentation issues in Figure 5 and its description. We reorganized Figure 5 and added more captions for each subfigure. Please see Section 4.2 to check the readability. We have made three changes: (1) add more captions for each figure in Section 4.2, especially for the y-axis. (2) We have re-examined the described patterns and revised the text to state explicitly which catchment exhibits each pattern. (3) we have replaced it with "saturation pattern" and revised the description to match what the figure shows in Sect. 4.2.

L540-542: But TDLR also uses Qt-1, doesn't it?

Response: Agreed. This has been clarified in Sect. 3.2 and Sect. 4.1. During TDLR training, Qt−1 is obtained from observations, while during continuous reconstruction, it can be replaced by the previously simulated Qt−1. We now explicitly report both the observed-Qt−1 reconstruction and the fully recursive/free-running TDLR experiment.

Table 2: Is this the reference for NSE?

Response: We thank the referee for pointing out the reference issue. We added the original NSE reference in the revised manuscript.

Table 2: what does "static/dynamic" mean here?

Response: Static and dynamic are two parameterization schemes of the dPL-HBV. The parameters of dPL-HBV are sixteen. The static scheme implies all parameters are time-invariant. The dynamic scheme implies that the beta and gamma parameters describing the infiltration and evapotranspiration processes are time-variant. We have added this explanation to the Table 2 caption and to Sect. 3.3 of the revised manuscript, following the original dPL-HBV formulation. For more details, please check Section 3.3 and the reference below.

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, e2022WR032404. https://doi.org/10.1029/2022WR032404

Figure 7: why are the lines in Figure 7 not the same as the ones in Figure 5, or are the differences just a scaling effect?

Response: The lines are the same. The apparent difference is due to the axis scaling. To prevent the misunderstanding, we added "Notably, the project curves are the same in Fig 8(a) and (b)." in the caption of Figure 9.

Reference:

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, e2022WR032404. https://doi.org/10.1029/2022WR032404

Gupta, H.V., Kling, H., Yilmaz, K.K. and Martinez, G.F. 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. Journal of hydrology 377(1-2), 80–91.

Knoben, W. J. M., Freer, J. E., and Woods, R. A. 2019. Technical note: Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores, Hydrol. Earth Syst. Sci., 23, 4323–4331,https://doi.org/10.5194/hess-23-4323-2019.

Nash, J., & Sutcliffe, J. 1970. River flow forecasting through conceptual models part I — A discussion of principles. *Journal of Hydrology*, *10*(3), 282-290. https://doi.org/10.1016/0022-1694(70)90255-6