Dear Editor,

We appreciate your patience in receiving these changes, as the first author now has changed institutions and has additional responsibilities. We appreciate the constructive feedback from you and both reviewers, and hope the following changes address your comments and concerns. Line numbers refer to the manuscript version with tracked changes.

Reviewer #1
The authors have indicated that this paper maintains the same model structure as a previous publication by the same research group. However, considerable effort has been invested in preprocessing the data and integrating it into the model pipeline, leading to the discovery of novel insights. Despite this, while the paper meets the criteria for publication and possesses scientific significance to a certain extent, it leans more towards generating insights from a new application rather than serving as a model evaluation paper. Consequently, I believe that this paper may find a better fit in a journal such as Water Resources Research, which aligns more closely with the findings presented. Therefore, I recommend that the editor reconsider whether to accept the paper in its current form or suggest that the authors submit it to a more suitable journal.

Thanks for your comments! GMD has a paper type called model evaluation which we think this paper fits quite well in. Many of the past papers published in this category evaluate an existing model structure under different contexts. As the reviewer acknowledged, a large amount of work was needed to run the model on the global scale.

Deep learning-based methods and the increasing amounts of observations provide great opportunities for global hydrologic modeling. In this paper, we evaluated both purely data-driven and physics-informed deep learning models globally on thousands of catchments. These models all support parameter regionalization, thus can be flexibly applied to predictions in ungauged regions over the world. Although spatial generalization is vital to global-scale modeling, few studies have systematically examined this issue with state-of-the-art deep learning and physics-informed deep learning methods at such a large scale as shown in this study. We also distill global hydrologic insights from the model evaluation to support future development and structure modification. This study is a necessary and important step to support the model selection for next-generation global models and perform highly accurate seamless modeling covering the global data-sparse regions. Therefore, we believe this study fits quite well to the scope of GMD for the model evaluation category.

Reviewer #2
General comments
Feng et al tested the performance of a hybrid model in simulating daily streamflow at several thousand global watersheds. They benchmarked the performance of the hybrid model with a deep learning model and found the hybrid model has a satisfied performance. The topic is a good fit to the scope of Geoscientific Model Development, though I agree with previous review that this study uses the same model and design of experiments from

the authors' previous study. The new insight of this work is to test the hybrid model framework in more watersheds locate from different continents. The idea of hybrid model is great and represents a significant contribution to hydrological modeling. It would be helpful for the authors to discuss the challenges of coupling the ML/DL model with a hydrological model. Specifically, since HBV model is relatively simple, is it possible to develop such a hybrid model with a more complicated hydrological model? Please also see my specific comments in the following.

Thanks for the comments!

We added the following paragraph to discuss how to implement a more complicated differentiable model and its challenges.

"To create a new differentiable model or turn an existing model into a differentiable one, we need to implement the model on a differentiable platform like PyTorch, Tensorflow, or JAX, while better enabling model parallelism in order to maximally leverage the computing power of modern graphical processing units (i.e., GPUs). If a model contains mostly explicit calculations, automatic differentiation (AD) offered by the above platforms can effortlessly provide gradient calculations, requiring only a syntax-level translation which can nowadays be done easily. Sometimes, a limited amount of adjustments are needed to turn non-differentiable operations into equivalent differentiable ones. However, when a model contains iterative solutions to nonlinear systems, large matrix solvers or constrained optimizations, we can employ the adjoint method (Song et al., 2023). The adjoint method explicitly defines the gradient-calculation method and alters the order of calculations so iteration is avoided during gradient calculations, which can dramatically reduce memory demand and improve efficiency. Another important consideration is the effective use of parallelism and the modern computing infrastructure for AI (i.e., GPUs). In our context, the regionalized parameterization (in this case, training one neural network on a large amount of basins), which is crucial to ensuring the generalizability of the model, requires going through large data in high-throughput parallelism. Embracing parallelism may necessitate some coding adjustments. At this point, several versions of differentiable hydrologic models have been proposed with varying complexities and different handling of parameterization, post-processing (which we didn't use in this study, as it can interfere with interpretability of the internal variables, mass balances, and the sensitivity to inputs encoded by the process-based components) and dynamical parameters. Across geoscientific domains, differentiable ecosystem (Aboelyazeed et al., 2023; Zhao et al., 2019), streamflow and river routing (Bindas et al., 2024), water quality (Rahmani et al., 2023), and ice sheet (Bolibar et al., 2023) models have already been demonstrated." {Lines 408-426}

This study mainly focuses on rainfall-runoff modeling, the key part of a more complicated hydrologic modeling system. We aim at identifying these two points from this study:

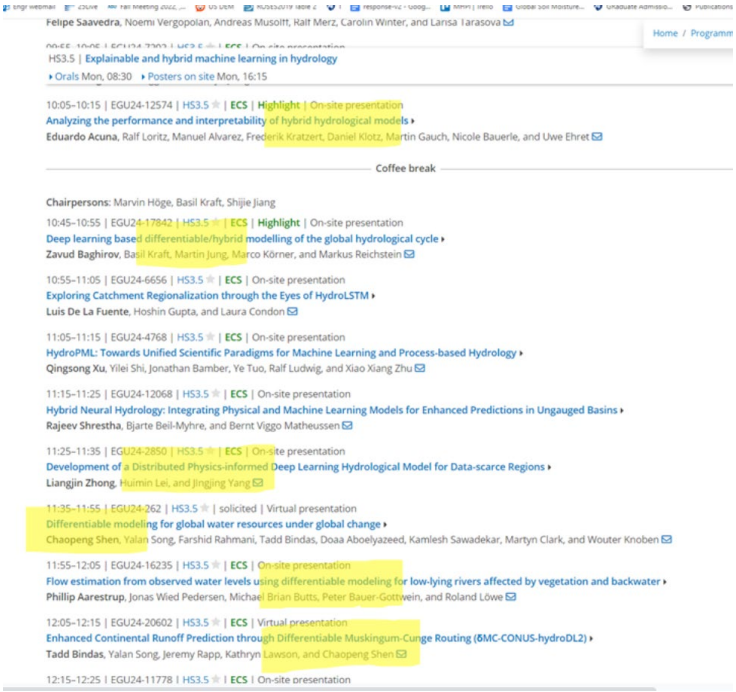
1) Which aspects of the used physical backbone should we improve next? Although the current framework provides strong performance advantages for large-scale modeling,

we identified several challenges from the evaluation process, such as the underrepresentation of processes like glaciers, ground ice, human interventions, etc. Future work should focus on improving these process representations in this modeling framework.

2) Does spatial generalization work for such a diverse global dataset with differentiable parameter regionalization? This is demonstrated by the spatial generalization tests which is very important to the global scale modeling given that many large lands across the world lack high-quality observations.

Beyond these two major aims, we can expect to integrate more physical processes in the future to develop the more complicated modeling system, as mentioned in the text: *"future work can gradually incorporate critical processes and include more observations to constrain the learning process, making sure each addition is valuable and accretive. The research community collectively has already substantial experience in evolving earth system models to include many processes. We expect some processes to be invited back in the differentiable modeling framework. Nevertheless, with differentiable modeling, we now have a new tool that was not previously available: highly flexible deep neural networks that can be placed anywhere in the model, which provide a systematic way of managing model complexity. With their help, such model evolution may take much less time than previously required. However, we still expect the development cycle to take longer than for purely data-driven models like LSTM, requiring us to view differentiable models as evolving rather than static entities, which need a bit of patience while maturing."* {Lines 453-461}

Several efforts of differentiable models have been developed or are under development as shown in the above added paragraph. We can potentially integrate these developments into the current framework to acquire more advanced modeling systems with better physical representations. Other groups in the community are also implementing complex land surface models on differentiable platforms. In the following two links (and screenshots), you can find other groups who are following up to do their differentiable models in hydrology or other fields.
https://meetingorganizer.copernicus.org/EGU24/session/48163



Specific comments

As the author argued in Line 47 – Line 55 that DL and LTSM have been demonstrated with good performance of simulating hydrological variables from local to continental scales, I wonder what is the novelty of this work? Although this study extends previous application at CONUS to global scales, the selected dataset doesn't have a good coverage for the whole globe. Please find my detailed comments regarding the selection of dataset in the Following.

We added these sentences to show the novelty of differentiable modeling to the LSTM model. *"Compared to the LSTM model which only outputs discharge simulations, differentiable models offer a suite of interpretable variables including ET, soil water, recharge, baseflow, etc., thus providing a comprehensive description for the hydrologic cycle and far better interpretability." {Lines 406-408}.*

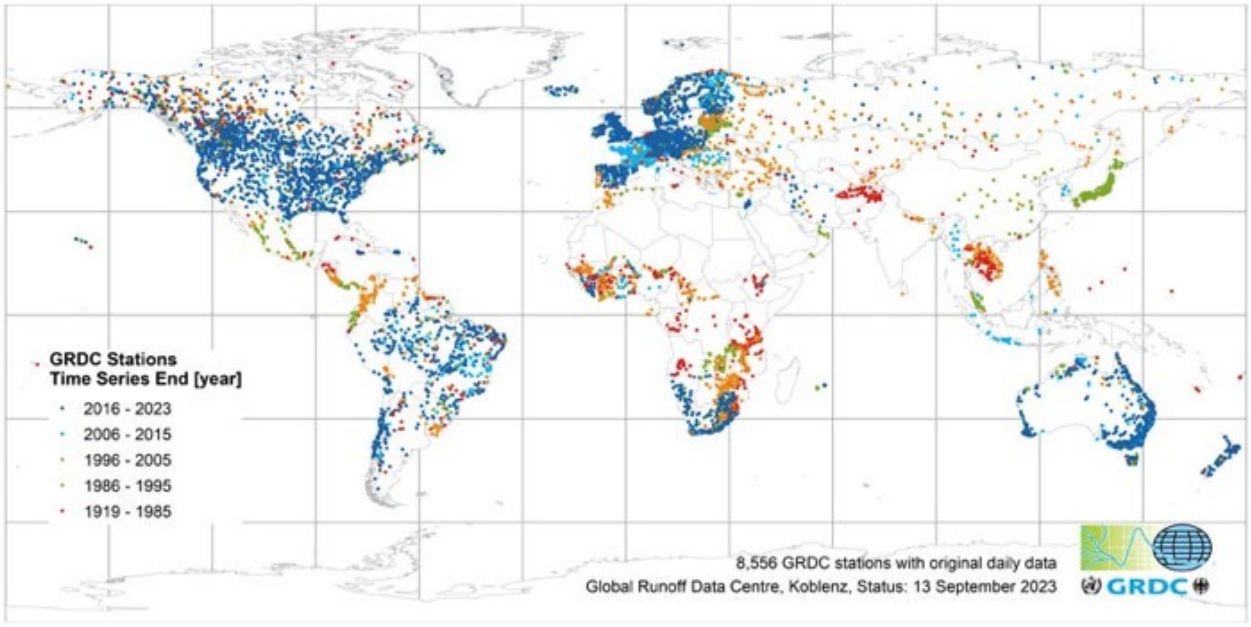
There are indeed some data issues that cause the gaps in the global map. However, the dataset already represents the best efforts and, in fact, even more comprehensive than some alternative contributions such as the Caravan dataset. The most original dataset from coauthor Hylke Beck's effort includes in total 21955 streamflow stations from several data sources. Then several criteria were applied to exclude some catchments including 1) without daily streamflow records; 2) "non-reference" catchments in GAGES-II with significant human interventions; 3) areas smaller than 50 km2 or larger than 5000 km2; 4) too short with less than 5 years data; 5) visual screening for erroneous streamflow data. Finally, a dataset with ~4000 catchments was formed with these criterions applied, as shown in Beck et al., 2020 that gives more details on the dataset formation. We further applied one round of manual screening on this dataset to

exclude some stations with apparently unrealistic records (like abrupt change and magnitudes of large differences in two periods) in our study.

As shown by our response in the last round, going from CONUS to global scale will require lots of work. We need real experiments in the global catchments to evaluate the applicability of different models at this large scale, especially for their spatial generalizability. All the global datasets have shown very sparse data availability in some continents like Africa and Asia, and how to improve the prediction in these regions is a key issue for global models and discussed by the extrapolation experiments in this study. In addition, we also gain lots of new insights from the large global dataset with diverse climate and geographic groups.

Line 124: Why the authors select this dataset given there exist other global streamflow datasets that contain much more gauges and have a better spatial coverage? Since the selected dataset archives headwater catchments, the observed streamflow is approximately same as the runoff generation, with the less river routing impacts. Is this because there is not river routing component in HBV? If runoff is the target variable, there are multiple well validated global runoff datasets to be used. Then it is not reasonable for the author to justify that the application is at global scales. In addition, Figure 1 shows the selected catchments are mainly from certain regions. Except North America, the gauges over other continents do not cover the continent uniformly, thus they are not representative for the continent.

Please see above our response to the last comment. We assume the reviewer is referring to a dataset like Caravan. Actually, while they are called different names, the core discharge data of most of these datasets were derived from the same data sources like GRDC (see the below Figure). You can already notice the gaps in Asia and Africa at a first glance. Also, many stations have stopped providing data and don't have long-term high quality data to support the modeling evaluation. Not all these stations and records are really available to use in the modeling work. As mentioned by the previous selection criteria, ~4000 catchments are the final selection amount after applying these necessary criteria. We will go to the similar catchments finally even if applying other compiled datasets like Caravan.



There is indeed internal routing in the model that could work up to about 5,000 km2 with effectiveness depending on the region. The reviewer is correct in that large river routing was not included --- this will be addressed in future studies. Since runoff is essentially not observable at large scales, current runoff datasets are mainly based on simulated products which will introduce additional bias to the evaluation. One of the main future works will be using the outcomes from this evaluation study to provide a global seamless runoff product. It's also worth noticing that, although the training is in basin-lumped format, this differentiable modeling has enabled learning the transfer functions to parameterize the model so that it can be easily applied to simulating all the terrestrial lands across the world. Differentiable routing is ongoing work (Bindas et al., 2024) and will be expanded to global scale in ensuing pursuits.

Line 95 – Line 97: There are several studies that calibrated the models to match monthly Variations.

We added this statement to include the literature that uses monthly discharge variation in calibration. *"Only very limited studies attempt to calibrate global models on monthly discharge variations (Werth and Güntner, 2010)."* {Lines 98-99}

Section 2.3: In traditional hydrological model calibration, one needs to run the physical model many times with perturbed parameters. How many forward simulations are needed in the hybrid model to identify the best parameter? Please clarify.

Deep learning-based models are trained in mini-batches for forwarding and optimization. Each mini-batch consists of a small subset group of training time sequences from the whole training data . Within one epoch we iterate and forward the model in each minibatch, and there are 1293 iterations in total. One epoch is defined as all the data points are forwarded once in a probability

concept after iterating over many mini-batches. We train our differentiable models for 50 epoches in total.
We added these statements in the method section to further clarify this point:
"Differentiable models are also trained in mini-batches that are formed in the same way as training the LSTM streamflow model. Within one epoch, differentiable models are forwarded and optimized over the randomly formed mini-batches until the iterations have used all the training data points. We train the differentiable models for 50 epochs in total." {Lines 185-187}

Line 202 – Line 218: I agree with the authors that PUR experiment is more challenging spatial extrapolation than PUB experiment. However, in practice, PUB experiment is more useful than PUR. Streamflow is probably the most well observed hydrological variables. At global scales, there exists abundant streamflow gauges with a good spatial coverage for each continent, though some continents have relative less than others. Therefore, one doesn't need to assume a whole continent is ungauged. It is the selected dataset in this study that gives sparse spatial coverage.

As mentioned in our response to the previous comments, it's not specifically related to the selection of our dataset. The reviewer can refer to the GRDC map shown above ---- lots of significant data gaps considering many gauges stopped reporting 20 years ago! If there were dense enough river gauge monitoring, there wouldn't have been satellite missions like SWOT --- which still won't fully fill the gaps due to its large-river nature. The alternative datasets share the same original sources and have the same issue that available streamflow records are quite sparse to use in some continents like Africa and Asia.

PUR is a general representation for the spatial extrapolation prediction, while PUB doing random hold-out is actually spatial interpolation, always having neighboring gauged basins as representative donors. Therefore, PUR is a more practical issue in the real world, because in most real cases, the whole large continuous regions miss observations, like the cases in Africa and Asia. The publicly available high-quality data are very sparse and we will often encounter the extrapolation issue when predicting in these large ungauged regions. However, deep learning based methods with high flexibility and transferability provide new opportunities to tackle these issues by learning from data-rich continents and transferring the learned information. Our cross-continent PUR experiments are just designed for this purpose.

Line 236: Do you mean a median KGE of 0.78?

Yes, thanks for pointing this out. We are referring to the median KGE value of the 1675 subset catchments with long-term streamflow observations. We have modified this sentence as "*LSTM even reached a median KGE of 0.78*" to clarify it. {Lines 239-240}

Section 3.2: I expect dPL + evolved HBV with DP is always better than dPL + evolved HBV, because the former model is more flexible to capture the observation. If no better, it should

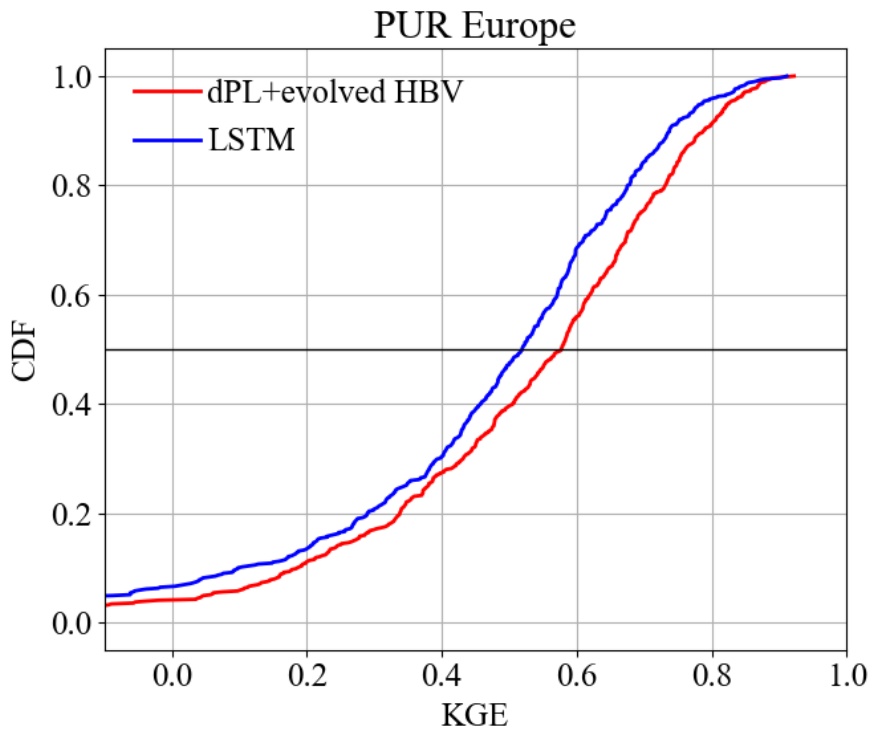
not be worse than the latter one. But Figure 5 shows, dPL + evolved HBV with DP is worse than dPL + evolved HBV in arid region. It will be helpful for the authors to clarify such Clarification.

The model with dynamic parameterization (dPL+evolved HBV with DP) is indeed more flexible than the other static parameterization model. That's why we always find the dynamic model has better fitting on the training data. However, Figure 5 is reporting the performance on the testing data that the models have never seen during the training. The dynamic model still performs better on the testing data in most cases, except for some specific cases like the temporal generalization in arid groups as the reviewer mentioned. These cases reflect another issue of overfitting. With more flexibility, the model also has a higher risk to be overfitted on the training dataset and reduce the generalization ability. This is why we only replace one or two parameters as dynamic ones to reduce this risk and also maintain physical clarity. In our previous paper we in fact warned against using too many dynamical parameters. In the CONUS dataset, the two-dynamic parameter model did not show more overfitting, presumably because the dataset was more homogeneous, while the behavior is somewhat different on this global dataset. This also shows that testing on a global dataset can give us different perspectives even on the similar issue. On the other hand, basins in the arid groups generally have lots of zero or very low flow rates, which makes these basins hard to simulate for all the models and very sensitive to the evaluation metrics. We use these statements in the main text to discuss these points:

"The differentiable model with dynamic parameters performed better than the model with static parameters in all climate groups except the most challenging arid group. Dynamic parameterization with more structural flexibility generally provides stronger modeling ability, while also showing a higher risk of overfitting and degraded generalizability in basins which are very difficult to simulate." {Lines 288-291}

Line 341: I don't think the median KGE = 0.58 is significantly better than median KGE = 0.52. They are pretty close performance to me. I suggest the authors to plot the cumulative density functions (CDFs) of the KGE for different model, and test if the CDFs are statistically different.

We performed the one-sided Wilcoxon signed-rank test and it shows that the KGE performance of the differentiable model is significantly better than the LSTM model (p-value less than 0.01). This can also be clearly seen from the box plot (Figure 6 in the main text) and the CDF plot below that the whole performance distribution of the differentiable model is leading that of the LSTM. We modified these statements to clarify this point, *"In the PUR scenario where European basins were held out for testing, differentiable models (median KGE=0.58) performed significantly better (p-value less than 0.01 using the one-sided Wilcoxon signed-rank test) than LSTM (median KGE=0.52)."* {Lines 348-350}



Line 350 – Line 351: Do you mean “cannot be obtained by straightforwardly training *physical* or *DL* models on data alone”? Figure 6 suggests the hybrid model is better than the traditional model calibration (Beck20). But it doesn’t support that performance of hybrid model is statistically better than a purely data-driven model.

The differentiable modeling indeed achieves better extrapolation performance for the cross-continent PUR predictions, as shown in our response to the last comment. We modified this statement as below to avoid confusion.

“*With these results, we show that differentiable models have demonstrated a high simulation capability that cannot be obtained with traditional parameter regionalization approaches, and robust extrapolation capability in large data-sparse regions that is stronger than purely data-driven models like LSTM*.” {Lines 358-360}

Line 371: Why not using KGE and NSE for the ET evaluation to be consistent with discharge evaluation?

The remote sensing ET products usually have bias errors dependent on different algorithms. Correlation and RMSE are more commonly used metrics when evaluating ET variables (Velpuri et al., 2013; Holmes et al., 2018), which reflect the consistency of temporal dynamics and the average magnitude of differences between two datasets. We keep using correlation and RMSE as used in the previous studies so that we can provide context for our simulations at global scale.

Holmes, T. R., Hain, C. R., Crow, W. T., Anderson, M. C., & Kustas, W. P. (2018). Microwave implementation of two-source energy balance approach for estimating evapotranspiration. *Hydrology and earth system sciences*, *22*(2), 1351-1369.
Velpuri, N. M., Senay, G. B., Singh, R. K., Bohms, S., & Verdin, J. P. (2013). A comprehensive evaluation of two MODIS evapotranspiration products over the conterminous United States: Using point and gridded FLUXNET and water balance ET. *Remote Sensing of Environment*, *139*, 35-49.